Étude de la composante génétique de l'efficience alimentaire (EA) chez des lignées de poules pondeuses divergentes pour l'EA en utilisant la technologie RNA-seq

Frédéric Jehl

DOCTORAT ECOLOGIE BRETAGNE GEOSCIENCES LOIRE AGRONOMIE ALIMENTATION

l'institut Agro agriculture • alimentation • environnement

AGRO CAMPUS OUEST

THESE DE DOCTORAT DE

L'INSTITUT NATIONAL D'ENSEIGNEMENT SUPERIEUR POUR L'AGRICULTURE, L'ALIMENTATION ET L'ENVIRONNEMENT
ECOLE INTERNE AGROCAMPUS OUEST

ECOLE DOCTORALE N° 600
Ecole doctorale Ecologie, Géosciences, Agronomie et Alimentation
Spécialité : *Génétique, génomique et bio-informatique*

Par

Frédéric JEHL

Étude de la composante génétique de l'efficience alimentaire (EA) chez des lignées de poules pondeuses divergentes pour l'EA en utilisant la technologie RNA-seq

Thèse présentée et soutenue à Rennes, le 16 décembre 2020
Unité de recherche : UMR 1348 PEGASE – INRAE / Institut Agro – équipe Génétique et Génomique
Thèse N° : 2020-31 C-150

Marie DE TAYRAC	MCU-PH, CHU Pontchaillou, Rennes
Gwenola TOSSER-KLOPP	Directrice de recherche, INRAE, UMR 1388 GenPhySE

Composition du Jury :

Présidente :	Maria MANZANARES-DAULEUX	Professeure, Institut Agro, UMR 1349 IGEPP
Examinateurs :	Marie DE TAYRAC	MCU-PH, CHU Pontchaillou, Rennes
	Gwenola TOSSER-KLOPP	Directrice de recherche, INRAE, UMR 1388 GenPhySE
	Jordi ESTELLE	Chargé de recherche, INRAE, UMR 1313 GABI
Directrice de thèse :	Sandrine LAGARRIGUE	Professeure, Institut Agro, UMR 1348 PEGASE
Co-directrice de thèse :	Tatiana ZERJAL	Chargée de recherche, INRAE, UMR 1313 GABI

たまごの近(ちか)くの 鶏(にわとり)

Poule près d'un œuf

食欲(しょくよく)は満(まん)たされた

Son appétit rassasié

魅惑(みわく)のバレエ

Fascinant ballet

Résumé de la thèse sous forme de haïku

Remerciements

Je tiens à adresser mes plus sincères remerciements à Sandrine Lagarrigue et Tatiana Zerjal pour leur encadrement et leur bienveillance durant cette thèse.
Sandrine, merci pour ton dynamisme communicatif, ta disponibilité de chaque instant et tes nombreux conseils, qui m'ont permis de passer trois belles et riches années.
Tatiana, merci pour tes conseils et tes relectures toujours rigoureuses, qui m'ont à chaque fois aidé à approfondir ma réflexion.
Merci à toutes les deux de m'avoir permis de saisir les opportunités qui s'ouvraient à moi et m'ont été précieuses : dispenser le cours de bio-informatique, encadrer des étudiants, réaliser une mobilité, participer à des congrès internationaux.
Un grand merci également à toutes les deux pour la relecture de la présente thèse.
También quiero agradecer a Yuliaxis Ramayo-Caldas, que tuvo la amabilidad de supervisarme durante mi estancia en el IRTA, pero también por su cálida acogida durante mi estancia.

Ensuite, mes remerciements vont à tous les membres de l'équipe que j'ai eu le plaisir de côtoyer.
Une pensée particulière pour Morgane, présente dans nos cœurs.
Colette, Morgane, Sophie, Laetitia, Pauline, Jean-Marc, Frédéric, Kévin, Florian, vous avez été chaque jour des collègues en or, toutes et tous ouverts et bienveillants.
Colette, merci pour ton énergie et ta gaieté rayonnante. Je te souhaite une belle retraite.
Sophie, merci pour ta bonne humeur, ton dynamisme et ta conscience professionnelle exemplaire.
Laetitia, merci pour les nombreuses validations expérimentales que tu as réalisées et ta bonne humeur de chaque instant.
Pauline, merci pour ton énergie et ton entrain communicatif.
Jean-Marc, merci pour ta présence et pour ta disponibilité.
Frédéric, merci pour ta grande diligence et pour la touche si personnelle que tu as apporté à ma thèse.
Kévin, tu as été un excellent co-thésard et je te souhaite une bonne continuation dans le monde de la recherche.
Florian, tu as toi aussi été un précieux compagnon de thèse, et je te souhaite également une bonne continuation dans la belle carrière qui t'attend.
Mes remerciements également aux membres de l'équipe installés à Saint-Gilles.

Fabien, nos quelques mois passés ensemble ont été un grand plaisir, et je te souhaite le meilleur pour ta thèse.
Merci aux membres de l'équipe *Genètica i millora Animal* de l'IRTA de Torre Marimon pour leur accueil, leur gentillesse et les conseils qu'ils m'ont dispensés lors de mon séjour parmi eux.
Merci aux membres de l'INRAE qui ont participé de près ou de loin aux travaux qui ont rendu possible le travail dans la présente thèse.
Merci également aux financeurs de cette thèse : la Région Bretagne et le département de Génétique Animale de l'INRAE.

Je tiens également à remercier les membres de mon comité de thèse, et particulièrement Andrea Rau et Thomas Derrien pour nos discussions et leurs conseils. De même, je remercie Yann Audic, Jean-Christophe Simon et David Renaudeau pour leur temps qu'ils ont accordé au comité.

J'adresse mes sincères remerciements à Marie de Tayrac, Gwenola Tosser-Klopp, Maria Manzanares-Dauleux et Jordi Estellé, qui ont bien voulu accepter d'être membres du jury de la présente thèse.

À mes amis en France et dans le monde, merci pour votre présence et votre soutien. *Gracias a Ana y Roma por su bienvenida en Caldas, y por soportar mi español básico y mi acento francés*.
Merci en particulier à Sophie, Julien, Marianne, Soledad et Philomène pour les belles années que nous avons passées à l'Agro et d'avance merci pour les belles années que nous passerons encore à l'avenir. Sophie, Marianne, Soledad et Julien, je vous souhaite une bonne thèse et une belle après-thèse. Sophie et Soledad, merci pour votre accueil à Montpellier. Merci Sophie pour nos discussions, et merci Soledad pour ta relecture de cette thèse. Philomène, je te souhaite une belle fin d'études et beaucoup de satisfaction comme vétérinaire.
Amis des soirées à distance du confinement, merci pour ces moments.
Merci bien-sûr à toute ma famille, et en particulier ma mère et ma grand-mère, pour leur soutien indéfectible et de tous les instants.

Merci enfin à toutes celles et tous ceux dont j'ai eu le plaisir de croiser la route et avec qui j'ai pu partager quelques instants.

Table des matières

La présente thèse est divisée en trois grands objectifs. **Le premier objectif** est d'étudier les tissus et gènes impliqués, voire responsables, de l'efficience alimentaire (EA), un caractère d'intérêt agronomique, chez la poule pondeuse, une espèce d'importance économique. Pour ce faire, nous avons utilisé comme modèle deux lignées de poules pondeuses très divergentes pour l'EA, obtenues après 40 ans de sélection sur un caractère appelé « prise alimentaire résiduelle ». Comme approche, nous avons comparé les transcriptomes de ces deux lignées dans quatre tissus : le tissu adipeux, le sang, l'hypothalamus et le foie, en faisant appel à une méthode de séquençage haut débit des ARN, le RNA-seq. Nous avons également étudié l'adaptation de ces deux lignées aux conditions d'élevage, en l'occurrence un régime hypo-énergétique, grâce à ces mêmes approches transcriptomiques pour éventuellement observer une interaction entre EA et régime alimentaire. **Le deuxième objectif** est de contribuer à l'annotation fonctionnelle du génome de la poule en utilisant toutes les données de RNA-seq générées dans le cadre du premier objectif. Nous avons tout d'abord enrichi l'annotation du génome de la poule en ARN longs non-codants, d'importants régulateurs de l'expression. Nous avons ensuite proposé une démarche et des critères que nous avons définis pour détecter systématiquement des variants de type SNP et les génotypes qui leurs sont associés à partir de données RNA-seq, données qui sont de plus en plus nombreuses et rarement utilisées à cette fin. **Le troisième objectif** enfin est d'étudier l'expression allèle-spécifique (ASE), qui combine à la fois SNP et expression, pour mettre en évidence des gènes candidats causaux de la variation d'efficience alimentaire à l'aide d'un dispositif particulier issu du croisement d'animaux de nos deux lignées et d'un outil récemment publié qui phase les SNP à l'échelle des gènes. Les résultats des deux premiers objectifs ont été utilisés dans le cadre du troisième.

Compte tenu de ces objectifs, l'introduction qui suit est composée de trois parties. Nous exposons d'abord l'importance d'étudier l'efficience alimentaire (EA) et l'adaptation à des conditions changeantes dans le contexte actuel de l'élevage, quelles que soit les espèces, puis présentons l'espèce poule en général et la filière pondeuse (donc les œufs) en particulier. Ensuite, nous développons le lien entre régulation de l'expression des gènes et caractères complexes (comme l'EA) d'une part et l'adaptation à des conditions changeantes d'autre part, justifiant ainsi la démarche mise en place. Enfin, nous présentons la méthode phare de cette thèse, le RNA-seq, et ses applications plus ou moins courantes, avant de proposer une synthèse des caractéristiques majeures des ARN longs non-codants, régulateurs importants mais encore méconnus de l'expression des gènes. Nous détaillerons ensuite nos objectifs, le modèle d'étude et le plan de la partie « Articles et travaux complémentaires ».

Introduction

I – Importance de l'efficience alimentaire et de la poule en élevage

Dans cette première partie, nous commençons par expliquer en quoi l'alimentation des animaux d'élevage représente un enjeu économique et environnemental important, sur lequel l'amélioration de l'efficience alimentaire permet d'agir, et en quoi les animaux d'élevage doivent parfois s'adapter à des variations de la qualité de leur ration (§ 1.). Ensuite, nous présentons l'espèce poule d'une manière générale, sous des aspects historiques, économiques et scientifiques, et d'une manière plus ciblée en nous intéressant aux œufs (§ 2.).

1. Efficience alimentaire et adaptation dans le contexte actuel de l'élevage

La production des aliments à destination des animaux d'élevage est responsable d'une large part des effets négatifs de l'élevage sur l'environnement et a un poids économique non-négligeable pour les filières (§ *a*). On souhaite donc travailler avec des animaux qui transforment une bonne partie de leur alimentation en produits d'intérêt économique : des animaux ayant une bonne efficience alimentaire (§ *b*). Mais les animaux d'élevage, en particulier les monogastriques comme la poule et le porc, sont issus de croisements entre des lignées sélectionnées dans des conditions maîtrisées, et sont exposés dans le temps et dans l'espace à des régimes alimentaires qui peuvent varier, notamment en qualité, et auxquels ils doivent s'adapter (§ *c*).

a) impact environnemental de la production des aliments destinés aux animaux

Le rapport sur la situation mondiale de l'alimentation et de l'agriculture publié par la FAO en 2009[1] distinguait dans son avant-propos trois grandes problématiques pour le secteur de l'élevage : (*i*) la pression qu'il exerce sur les écosystèmes, les ressources naturelles et sa contribution au changement climatique, point que nous développerons dans la suite, (*ii*) les risques sanitaires que la mondialisation du secteur fait courir aussi bien aux animaux qu'aux humains, et enfin (*iii*) la manière dont les productions animales contribuent ou non à la réduction de la pauvreté.

Une part de l'impact de l'élevage sur l'environnement est liée non pas directement aux animaux, mais à l'alimentation qui leur est nécessaire. En effet, si 14.5% des émissions de gaz à effet de serre d'origine humaine sont dues à l'élevage (le plaçant en 4^{è} position dans les secteurs émetteurs de gaz à effet de serre, derrière la production d'énergie, l'industrie, la

déforestation, et devant le transport[2]), 45% d'entre elles sont liées à la production, la transformation et le transport des aliments à destination des animaux[3]. Les gaz en question sont majoritairement du CH_4 (méthane), du N_2O (protoxyde d'azote), et du CO_2 (dioxyde de carbone)[4]. La production d'aliments à destination des animaux représente également 33% de l'utilisation des terres arables dans le monde, et aux États-Unis, 37% de l'utilisation des pesticides. À cela on peut ajouter que 30% des surfaces déforestées en Amazonie sont destinées à ce type de cultures, avec les pertes de biodiversité associées. Le secteur de la volaille est particulièrement concerné, puisqu'en 2016, la fabrication d'aliments composés pour volailles représentait environ 43% du volume total d'aliments composés fabriqués pour animaux d'élevage[5]. Notons en revanche que les productions de viande et d'œufs de poules sont les plus faibles émettrices de gaz à effet de serre par kg de protéines : en moyenne, respectivement 40 et 42 kg eq. CO_2 / kg protéine, contre 52 kg eq. CO_2 / kg protéine pour le porc, 125 à 189 kg eq. CO_2 / kg protéine pour le lait et la viande de petits ruminants, 84 kg eq. CO_2 / kg protéine pour les bovins lait et 342 kg eq. CO_2 / kg protéine pour les bovins viandes[6].

En plus de ces conséquences sur l'environnement, l'alimentation destinée aux animaux d'élevage a un poids important dans les revenus du secteur. En effet, on estime que l'aliment représente entre 60% et 70% des coûts de production, aussi bien chez les ruminants[7] que chez les monogastriques (porcs[8], volailles[9], dont poules pondeuses[10]).

Si l'élevage a des impacts négatifs sur l'environnement, notamment, comme on l'a vu, par la production d'aliments destinés aux animaux, il n'en fournit pas moins une large gamme de biens et services environnementaux, économiques et sociaux[11] qu'il convient de ne pas négliger. En suivant Ryschawy *et al.* (2015)[12], nous pouvons distinguer quatre grands ensembles de biens et services. D'abord, l'ensemble « approvisionnement », sous-entendu en produits et en coproduits de l'élevage, qui parait bien sûr en être l'objectif premier. Ensuite, l'ensemble « vitalité territoriale », qui passe par la création d'emplois ou le maintien du tissu rural par exemple. L'ensemble « qualité environnementale » passe notamment par la séquestration du carbone dans les prairies, la contribution au cycle des nutriments ou le maintien de la biodiversité dans les prairies[13]. Enfin l'ensemble « Patrimoine et qualité de vie », passe par exemple par le maintien de paysages végétaux et animaux diversifiés (exemples : bocage normand *versus* estives alpines pour les végétaux ; races Parthenaise *versus* Salers pour les animaux) ou par le patrimoine gastronomique (labels de qualité, gastronomie régionale). L'importance de ces différents biens et services et leurs natures exactes dépendent bien entendu du type d'élevage et de la zone géographique concernés.

b) l'efficience alimentaire, ou la valorisation de l'énergie de l'alimentation par l'animal

Conceptuellement, l'efficience alimentaire (en anglais, *feed efficiency*) décrit la manière dont l'animal valorise l'aliment pour une ou des fins données, qui sont, dans les cas qui nous intéressent, des fins de production. L'efficience alimentaire est un caractère qui semble être discuté dans la littérature anglophone depuis le début des années 1930 (recherche Google Ngram des termes « *feed efficiency* » et « *feed utilization* »), et PubMed date la première mention de ces termes dans les années 1940, avec notamment une étude s'intéressant à l'efficience chez les bovins, les ovins, les porcs et les volailles[14,15]. Concomitamment, les premières études sur l'héritabilité et la sélection[16] pour ce caractère sont publiées, notamment chez le poulet de chair[17], la souris[18] ou le rat[19], ouvrant ainsi la porte à son amélioration dans les espèces d'élevage par sélection génétique, puis par sélection génomique ces dernières années.

L'efficience alimentaire peut se quantifier grâce à l'indice de consommation (*Feed Conversion Ratio*, FCR) qui consiste en la division de la masse d'aliment ingéré par la masse de produit généré par l'animal[20] (viande, lait, œufs) :

$$\boxed{\mathrm{FCR} = \frac{\text{masse d'aliments ingérés}}{\text{masse de produits}}}$$

Une autre manière de quantifier l'efficience alimentaire est de calculer la prise alimentaire résiduelle (*Residual Feed Intake*, RFI)[21]. La RFI consiste en la différence entre la prise alimentaire observée et la prise alimentaire prédite par un modèle de régression linéaire multiple avec pour variables explicatives les besoins de production, de maintien, et, lorsque ceux-ci ne sont pas confondus avec la production (cas des poules pondeuses et des vaches laitières), les besoins de croissance. On a donc :

$$\boxed{\mathrm{FI}_{\text{prédite}} = a\mathrm{BW}^{0.5} + b\Delta\mathrm{BW} + c\mathrm{E}}$$

Où $\mathrm{FI}_{\text{prédite}}$ est la prise alimentaire prédite, BW est le poids du corps, ΔBW la variation du poids du corps sur la période et E la masse d'œufs produite, et :

$$\boxed{\mathrm{RFI} = \mathrm{FI}_{\text{observée}} - \mathrm{FI}_{\text{prédite}}}$$

On peut donc voir la RFI comme la part de la prise alimentaire non destinée au maintien, à la croissance ou à la production. On voit donc qu'au contraire du FCR, le RFI est indépendant des besoins de production, et constitue donc une autre mesure de l'utilisation de l'énergie alimentaire.

L'efficience alimentaire varie d'une espèce à l'autre, mais également au sein même d'une espèce, en fonction du système d'élevage, de la nourriture ou encore de l'âge et du poids de l'animal[22]. Des FCR trouvés dans la littérature pour différentes espèces d'élevage sont présentés Tableau 1. Si l'on excepte les bovins laitiers, qui produisent du lait, composé à 90% d'eau[23], ce qui explique que le FCR soit inférieur à 1, c'est le poulet de chair, suivi par les poules pondeuses, qui ont les FCR les plus faibles parmi les espèces d'élevage terrestres (1.85 à 2.10). Les ovins et bovins viandes, espèces ruminantes, ont quant à eux les FCR les plus élevés (supérieurs à 5). L'aquaculture semble présenter la meilleure efficience alimentaire, mais la production aquacole mondiale représentait environ 70 millions de tonnes en 2012[24], à comparer avec 66 millions de tonnes de viande bovine produites en 2015, qui est la plus faible production comparée aux œufs et aux viandes de volailles et de porc (voir aussi Figure 2, page 13). L'entomoculture (élevage d'insectes) enfin, présente également une très bonne efficience alimentaire, mais la production européenne d'insectes (les chiffres mondiaux pour les productions d'insectes sont plus difficiles à trouver) est anecdotique, avec environ 6000 tonnes par an, surtout destinées à l'alimentation animale[25].

Tableau 1 | **FCR moyens de différentes espèces**. Des extrêmes trouvés dans la littérature sont indiqués entre crochets.

Espèce	**FCR indicatif**[€]	**[extrêmes]**	**Références FCR**	**Production**[£] (tonnes)
Poule pondeuse	2.10	[1.97 – 2.30]	[26–28]	77 000 000
Poulets de chair	1.85	[1.70 – 2.00]	[26,29]	114 000 000
Autres volailles[¥]	2.60	[2.40 – 2.80]	[30,31]	13 000 000
Porc	3.20	[2.80 – 3.60]	[26,29,32]	121 000 000
Bovin laitier	0.75	[0.55 – 0.90]	[28,33]	683 000 000
Bovin viande	6.25	[6.00 – 6.50]	[29,34]	67 000 000
Ovins viande	5.10	[4.10 – 6.10]	[35,36]	10 000 000
Aquaculture	1.40	[1.00 – 1.80]	[29,37,38]	70 000 000
Entomoculture	1.80	[1.50 – 2.10]	[39,40]	*6000*

[€] kg matière sèche / kg produit. « Produit » est le lait ou les œufs (pour « bovin laitier » et « poule pondeuse »), le poids vif de l'animal pour le reste

[£] Production mondiale, d'après [41] sauf pour l'aquaculture ([24]). Pour l'entomoculture, production européenne d'après [25]. Voir aussi Figure 2 plus loin. Masse d'œufs pour « poule pondeuse », masse de lait pour « bovin laitier », masse de viande ou assimilé pour le reste.

[¥] Dinde, pintade et canard de barbarie

c) la nécessaire adaptation des animaux d'élevage à leur environnement

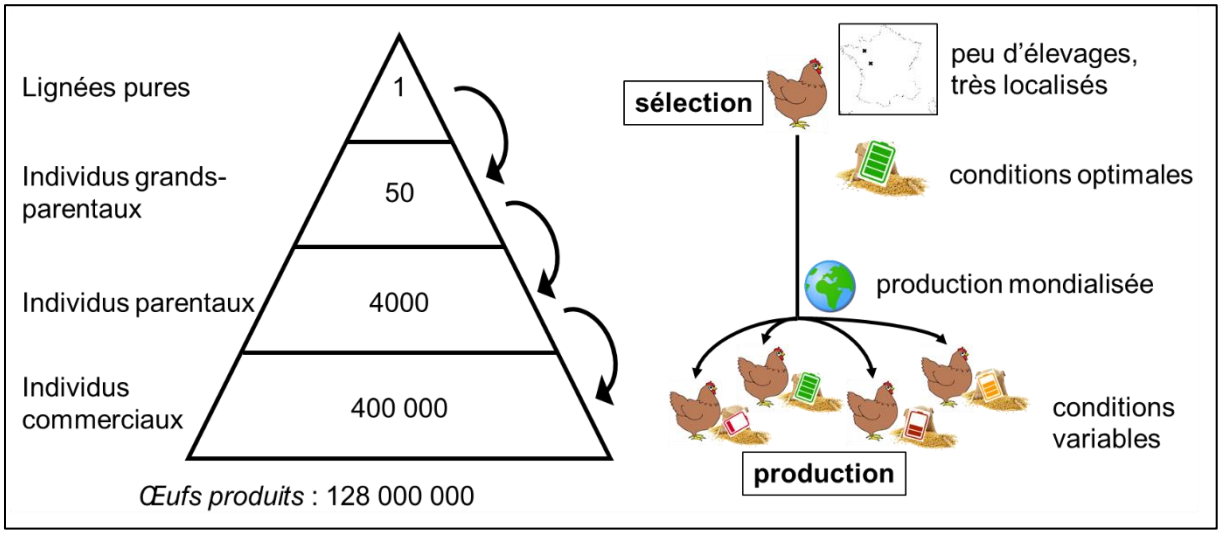


Figure 1 | Organisation pyramidale de la filière ponte. *Adapté de* [42]. Les nombres dans la pyramide représentent proportionnellement le nombre d'individus de chaque étage pour la filière poule pondeuse de l'entreprise Novogen (France). *Gauche* : illustration de schéma « pyramidal ». Les animaux des trois ou quatre lignées pures sont croisés entre eux pour donner les individus grands-parentaux, donnant eux-mêmes par croisement les parentaux, et finalement les commerciaux. Un animal de la lignée pure contribue à la production de 400 000 individus commerciaux, et ils produisent eux-mêmes 128 millions d'œufs. *Droite* : illustration de deux différences majeures entre les étages de sélection et de production. Les premiers ont lieu dans peu d'élevages dont les conditions d'élevages sont contrôlées et optimales (ici, le contenu énergétique de la ration), alors que la production se fait partout dans le monde et dans des conditions variables, possiblement sous-optimales.

La diffusion du progrès génétique chez les monogastriques (poules, porcs) se fait, contrairement aux ruminants, selon un schéma dit pyramidal, en raison du nombre important de descendants que peut produire un parent (Figure 1). Les élevages de production (qui abritent des animaux « spécialisés » : poules pondeuses et poulets de chair), très nombreux, présents partout dans le monde et gérés par de nombreux éleveurs, sont physiquement différents des élevages de sélection, peu nombreux, concentrés dans quelques lieux et gérés par quelques entreprises. Ces derniers élevages, au sommet de cette pyramide, abritent des animaux de trois ou quatre lignées pures, qui sont croisés entre eux pour donner les individus « grands-parentaux ». Ces animaux sont croisés entre eux, donnant naissance aux reproducteurs (« parentaux ») qui constituent l'étage de multiplication, en dessous de l'étage de sélection. Enfin, ces reproducteurs sont à nouveau croisés entre eux, donnant naissance aux animaux de l'étage de production (ce sont donc les petits-enfants des animaux de l'étage de sélection), dont les produits sont utilisés dans le commerce et bénéficient du progrès génétique de la génération de leurs grands-parents (génération $N - 2$). Dans cette structure pyramidale, les petits-enfants des animaux de l'étage de sélection sont confrontés à des conditions d'élevages très diverses, aussi bien dans l'espace que dans le temps. Ainsi, la ration alimentaire des poules est composée à plus de 60% de maïs

en Chine, *versus* moins de 20% en France, et des difficultés d'approvisionnement liées au contexte économique ou politique peuvent faire varier la composition des rations. Dans le cas des poulets de chair et poules pondeuses, les quelques entreprises de sélection qui se partagent le marché mondial (pour les pondeuses : Novogen, Française ; Erich Wesjohann Group, Allemande et Hendrix Genetics, Néerlandaise ; pour les poulets de chair : Cobb Vantress, Américaine ; Aviagen, qui appartient à Erich Wesjohann Group) n'ont évidemment pas de lignées sélectionnées pour chaque condition de production. Ainsi, les animaux doivent s'adapter à ces conditions variables.

2. *Gallus gallus domesticus* : histoire et importance économique, scientifique et sociale de la poule

Les travaux entrepris dans la présente thèse l'ont été sur l'espèce poule, particulièrement sur la poule *pondeuse* pour toute la partie liée au contexte agronomique. Comme le montrera notre exposé sur l'histoire de l'espèce depuis sa domestication (§ *a*), la nuance « ponte *versus* chair » n'avait pas vraiment de sens avant le XIX^{e} siècle. Nous présenterons l'intérêt économique des deux filières (§ *b*), avant de faire un bref focus sur l'œuf, et en particulier sur son intérêt nutritionnel (§ *c*). Enfin, faut-il le préciser, la poule est une espèce utilisée dans différents domaines de la recherche. Nous terminerons cette partie en citant quelques champs de recherche dans laquelle elle est utilisée, et en présentant quelques caractéristiques de son génome (§ *d*).

a) une brève histoire de la poule : de la domestication aux lignées commerciales

On dénombre actuellement 5 représentants du genre *Gallus* : la poule de jungle (*G. gallus*, ou *Red Jungle Fowl*) qui est l'ancêtre des poules domestiques actuelles (*G. gallus domesticus*), le coq de Sonnerat (*G. sonneratii*), que l'on trouve en Inde, le coq de Ceylan (*G. lafayetii*) que l'on trouve au Sri Lanka et le coq de Java (*G. varius*) que l'on trouve en Indonésie[43]. On estime que la poule a été domestiquée il y a environ 7 000 – 10 000 ans et à différentes reprises en Asie du Sud-Est et en Chine[44]. Dans les premiers temps suivant sa domestication, la poule semble avoir eu un rôle principalement, voire exclusivement, social et culturel plutôt qu'alimentaire. Ainsi, des représentations de combat de coqs ont été retrouvées sur le site archéologique de Mohenjo-Daro, dans l'actuel Pakistan, qui aurait été occupé par la civilisation harappéenne entre 2500 et 1800-1900 avant le présent, et des textes chinois de la dynastie Shang (1765 – 1122 av. J-C.) évoquent son rôle sacrificiel[45]. Par la suite, la poule serait arrivée en Europe en passant par la Russie ou par la Perse, avant d'arriver en Méditerranée et notamment

en Grèce autour de 700 av. J-C[46]. Là encore, il semble que son usage ait été plus récréatif qu'alimentaire, en Grèce comme dans tout l'Ouest de l'Europe actuelle : Jules César (102 ou 100 – 44 av. J-C.) évoque par exemple dans ses *Commentaires sur la Guerre des Gaules* le fait que les habitants de l'actuelle Grande-Bretagne ne consommaient ni lièvre, ni poule, ni oie, par interdit alimentaire, mais les élevaient « pour le plaisir » (Livre V, 12)[47], c'est-à-dire probablement pour les combats. Les Romains justement ont développé l'élevage à des fins alimentaires en plus des fins récréatives, comme l'évoque l'agronome Columelle (4 – 70 ap. J-C.) dans son *De re rustica* : « *Le revenu qui provient de ces oiseaux de basse-cour* [les poules] *n'est pas à dédaigner* […] *nous ne partageons pas le goût des Grecs, qui élevaient le coq, ce fier oiseau, pour les joutes et le combat* »[48]. Ils auraient notamment créé des races spécialisées et raisonnablement productives, mais l'Empire dans sa chute a entrainé avec lui ce secteur[49]. Les Romains utilisaient également des poules à des fins divinatoires. Ainsi, durant les campagnes militaires, des poulets sacrés dont s'occupait un soldat spécialisé, le *pullarius*[50], étaient utilisés : la manière dont ils mangeaient – ou non – la nourriture qui leur était présentée actait l'approbation divine[51]. L'introduction de la poule en Afrique et en Amérique du Sud est moins bien documentée. Pour l'Afrique en général, les plus anciens ossements ont été découverts en Égypte et datent de la dix-huitième dynastie (env. 1567 – 1320 av. J-C.). Pour l'Afrique sub-saharienne en revanche, il existe très peu d'éléments datant d'avant l'an mil[52,53]. En ce qui concerne l'Amérique du Sud, il est possible que les poules aient été introduites depuis l'Océanie à l'époque précolombienne, autour de 1300 – 1420 ap. J-C[54].

Les œufs ont pour leur part eu une importance symbolique majeure dans la cosmogonie de nombreuses civilisations. Il semble admis que cette importance est liée à la symbolique de « l'apparition » de la vie portée par l'œuf. On retrouve également l'œuf comme un des éléments de la fête chrétienne de la Pâques, et il a d'ailleurs à cette occasion été sublimé par Fabergé, sur commande du tsar Alexandre III, par la création des « œufs de Fabergé ».

La domestication et l'élevage ont entrainé l'apparition de races ayant des caractéristiques propres. Il est intéressant de constater que ces races domestiquées ont en commun un certain nombre de traces de sélection, qui sont des régions du génome dans lesquelles les différences de fréquences alléliques avec les lignées sauvages ne sont pas dues qu'à la dérive génétique, c'est-à-dire au hasard, mais à la sélection par l'être humain. L'une de ces traces contient notamment le gène *TSHR* (*Thyroid stimulating hormone receptor*)[55], qui joue un rôle dans le métabolisme et le comportement, en particulier la reproduction et la sociabilité[56]. Plus précisément, il semble qu'un polymorphisme G > A (rs13587540) dans *TSHR*, entrainant une

mutation faux sens Glycine > Arginine, et dont la fixation a eu lieu au court de 500 dernières années[57], puisse avoir contribué à la domestication[58].

Au cours du Moyen-Âge, les races poursuivent leur évolution, et œufs et viande de poule sont consommés sans restrictions particulières, comme l'attestent les recettes de l'époque[59]. Les poules ont alors une importance économique assez négligeable, et on en trouve entre moins d'une dizaine à quelques dizaines dans les basse-cours, servant parfois de monnaie d'échange[49]. C'est par ailleurs du Moyen-Âge que pourrait dater l'association symbolique du coq et de la France. Si l'origine de cette association est incertaine (un jeu de mot latin entre « *Gallus* », le gaulois et « *gallus* » le coq, la poule ?), le coq fait néanmoins aujourd'hui partie des symboles de la France[60]. Il se retrouve notamment sur le sceau de l'État, différentes pièces de monnaies anciennes (par exemple les pièces de franc-or à partir de 1899), comme emblème sportif, et au sommet de nombreux monuments aux morts de la première guerre mondiale, durant laquelle son image a été très utilisée pour personnifier la France. C'est au XIX^{e} siècle que la production d'œufs et de viande entre dans l'ère industrielle, avec la création d'entreprises de production voire de clubs de « sélectionneurs » en Angleterre (*Sussex Club, Utility Poultry Club*), et on date de la fin du XIX^{e} ou du début du XX^{e} siècle la divergence entre poulets de chair et poules pondeuses[55]. À partir de ce moment, les progrès dus à la sélection génétique, puis très récemment à la sélection génomique, ont permis de créer des lignées de poules pondeuses produisant autour de 300 œufs de couleur homogène par poule et par an et des lignées de poulets de chair capable d'atteindre 2 kg à l'âge d'abattage de 6 semaines.

b) importance économique des productions de viande et d'œufs

En termes de masse de produits, et en comparaison avec les bovins et le porc, le poulet se classait en 2016 deuxième derrière le porc (121 millions de tonnes pour le porc *versus* 114 millions pour le poulet), et les œufs devancent les bovins pour la troisième position (77 millions de tonnes pour les œufs *versus* 67 millions pour les bovins, voir aussi Figure 2A). En revanche, en valeur, viande et œuf sont troisième et quatrième, respectivement (Figure 2B). Le fait que la masse de viande de poulet produite en 2016 soit proche de celle de viande de porc implique une disparité dans le nombre d'animaux abattus, vu les poids respectifs d'un poulet (autour de 2 kg) et d'un porc (100-150 kg). En effet, en 2018, 70 milliards de poulets ont été abattus (leurs carcasses devenant le « fossile de l'anthropocène »[61], tant les carcasses s'accumulent dans les décharges mondiales), contre 1.5 milliards de porcs, 300 millions de bovins, et 7.5 milliards de poules pondeuses étaient utilisées pour la production d'œufs[41,62] (Figure 2C).

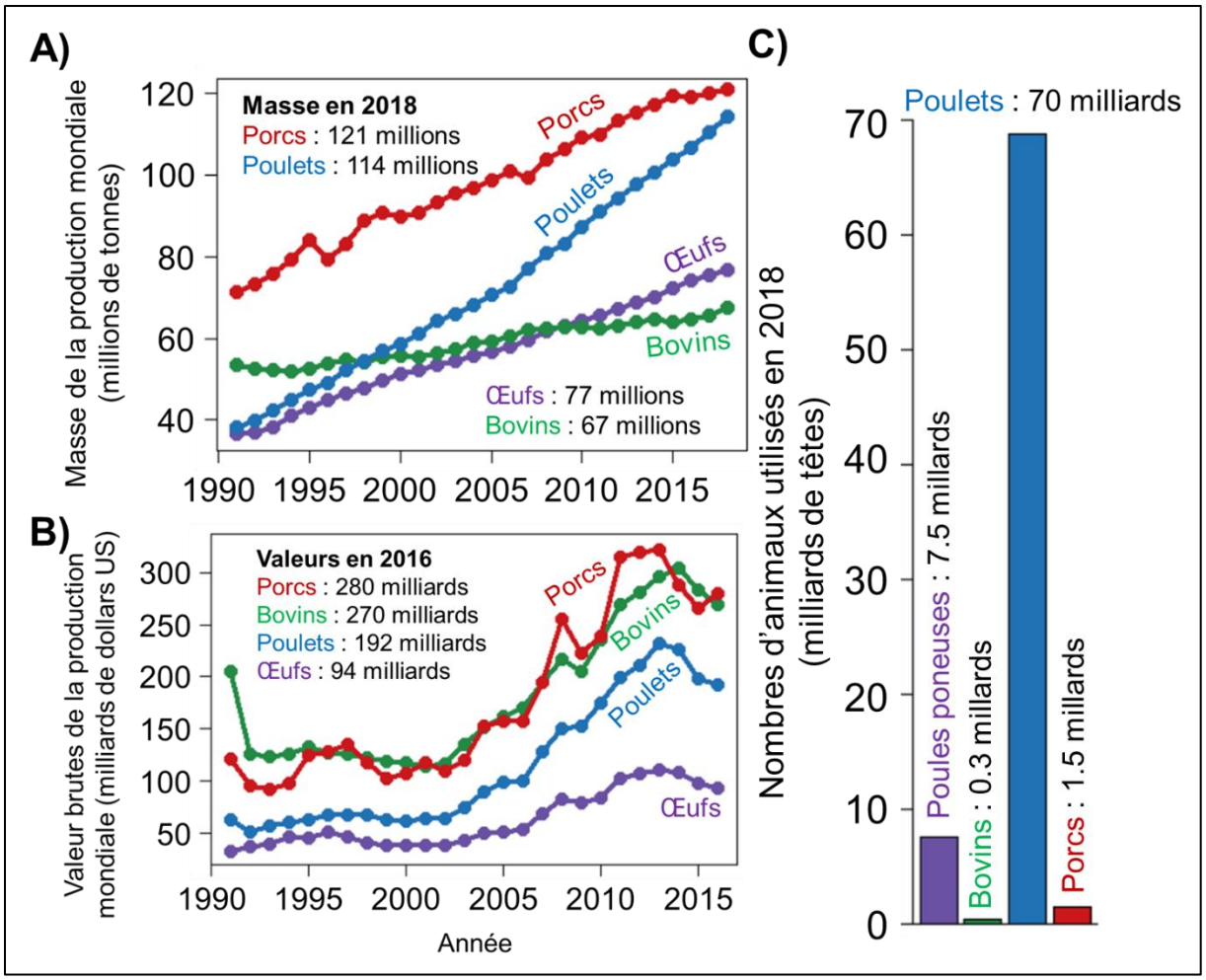


Figure 2 | Évolution des productions de viandes de porcs, poulets, bovins ainsi que des oeufs de poules en masse (millions de tonnes, **A**) et en valeur (milliards de dollars US, **B**) entre 1991 et 2018 (**A**) ou 2016 (**B**). Nombre d'animaux abattus ou utilisés pour la production en 2018 (milliards de têtes, **C**).

En 2018, le continent américain produisait environ 42% de la viande de poulet (en masse) à l'échelle mondiale, suivi de l'Asie (35%) et de l'Europe (16%). La production africaine représentait quant à elle 5% de la production mondiale, et le solde (1% environ) revenait à l'Océanie. Les États-Unis, le Brésil et la Chine se partageaient le podium, avec respectivement 17.5%, 13% et 12% de la production mondiale, la France occupant la 27^{e} place avec 1.2 millions de tonnes, soit 1.05% de la production. Pour les œufs en revanche, la production est beaucoup plus concentrée. En effet, toujours en 2018, l'Asie produisait environ 60% des œufs dans le monde, suivie par le continent américain (21%) et le continent européen (14%), le solde (environ 5%) venant d'Afrique et d'Océanie. Le Chine occupait la première place mondiale, produisant à elle seule 27 millions de tonnes d'œufs, soit 35% de la production mondiale, suivie par les États-Unis (6.5 millions de tonnes, 8% de la production). La France quant à elle occupe

la 12è place, avec une production d'environ 850 000 tonnes (1.1% de la production mondiale), ce qui en fait le premier pays de l'Union Européenne pour cette production[41,63].

À l'horizon 2050, les tendances observées sur la Figure 2A devraient se poursuivre, la production de viande de poulet dépassant celle de viande de porc (181 millions de tonnes *versus* 143 millions de tonnes), et les productions d'œufs et de viande de bœuf restant proche l'une de l'autre (102 millions de tonnes pour les œufs, 106 millions de tonnes pour la viande de bœuf)[64].

En plus de cette importance économique et industrielle, l'élevage rural de poules et poulets fournit jusqu'à 20% des apports de protéines animales dans la plupart des pays africains[65]. Ces animaux, qui font généralement partie de systèmes extensifs dans lesquels ils cherchent eux-mêmes leur nourriture ne sont pas particulièrement sélectionnés et représentent 80% des 800 millions de poulets du continent[66]. Cette forme d'élevage ne permet qu'une faible production, mais implique des investissements limités et est donc peu risquée financièrement. En plus des rôles sociaux et traditionnels des animaux[67], on considère que cette forme d'élevage, dont la gestion est généralement dévolue aux femmes et aux jeunes permet leur autonomisation, en particulier dans les zones rurales. Enfin et surtout, l'élevage de poules et poulets fournit des aliments riches en protéines et nutriments, qui sont en plus, dans le cas des œufs, stockables dans les conditions locales.

c) l'œuf, « nature's perfect food »

L'œuf de poule, en tant que cellule reproductrice, doit pouvoir assurer la survie et le développement du poussin[68]. Pour cela, il doit pouvoir fournir l'ensemble des nutriments nécessaires (stockés en particulier dans le jaune) et assurer une défense immunitaire (protéines immunitaires du blanc). Cela en fait un aliment de bonne qualité pour d'autres espèces animales comme les humains, et lui procure même le titre de « *nature's perfect food* »[69], de la part de certains auteurs enthousiastes.

La formation de l'œuf commence dans l'ovaire de la poule, avec l'accumulation de lipides et de protéines dans le cytoplasme de l'ovocyte, qui donnera le jaune d'œuf. Cette accumulation commence 4 à 6 jours avant l'ovulation, à raison de 2 g par jour[70]. À part les immunoglobulines issues du sang (IgY), toutes les protéines et lipides du jaune ont pour origine le foie de l'animal, qui joue donc un rôle majeur chez les poules pondeuses. L'ovulation a lieu environ 24h avant la ponte. L'ovocyte expulsé par cette ovulation entre dans l'oviducte, dans lequel il subira les dépôts successifs des différentes composantes de l'œuf. D'abord, la membrane vitelline (10-12 µm d'épaisseur), qui limite les contacts entre le jaune et le blanc, puis les protéines de l'albumen (c'est-à-dire le blanc), qui sont hydratées un peu plus loin. Après le dépôt du périalbumen, qui

entoure l'albumen, et des deux membranes coquillères (interne puis externe), l'œuf entre dans l'utérus dans lequel il restera pendant 20h. Il y subira pendant 12h le dépôt de la coquille[71,72]. Enfin, 2h avant la ponte, l'œuf est entouré par la cuticule. Ce processus recommence entre 15 et 45 minutes après la ponte[71].

Dans un œuf frais, le blanc d'œuf représente 60% de la masse, le jaune 30% et la coquille 10%. Le blanc est composé à 90% d'eau, et à 10% de protéines et d'une faible quantité de glucose[71]. Parmi les protéines du blanc d'œuf, les protéines majoritaires sont les ovalbumines, dont les rôles biologiques ne sont pas encore bien compris. Ce sont en revanche ces protéines qui sont responsables de la coagulation du blanc à la chaleur. De manière générale, le blanc d'œuf a d'importantes propriétés antibactériennes, en particulier pendant la première moitié des 21 jours que dure l'incubation du poussin[73]. Le jaune d'œuf est lui composé à 50% d'eau. La majeure partie de la matière sèche du jaune d'œuf est composée quant à elle de triglycérides et de cholestérol (63%), le reste étant des protéines (33%), des sucres et des minéraux (~4%)[74].

Ainsi, l'œuf est considéré comme une très bonne source de protéines, dont les proportions en acides-aminés sont particulièrement adaptées aux besoins de l'être humain, et de graisses facilement digestibles. C'est en outre une bonne source de micronutriments (phosphore, fer et certaines vitamines). Il est en revanche pauvre en glucides, en calcium (pour ce qui est bien sûr de la partie consommable de l'œuf) et en vitamine C[71].

d) une espèce modèle à bien des égards

La poule a plusieurs usages en recherche, en plus d'apparaitre dans des titres d'articles souhaitant exprimer le problème de causalité « qui de la poule ou de l'œuf ? » (c'est l'œuf[75]). C'est un animal de petite taille, facile à entretenir et avec un temps de génération court (moins d'un an). Cette espèce est utilisée pour la recherche sur le développement des vertébrés puisque les embryons contenus dans les œufs sont facilement accessibles et manipulables[76], mais également en virologie (première découverte des rétrovirus[77]) ou encore en immunologie (c'est de la bourse de Fabricius, organe lymphoïde propre aux oiseaux, que les lymphocytes B tirent leur nom[77,78]). De plus, le groupe des *Aves* (auquel appartient la poule) et l'être humain ayant divergé il y a environ 300 millions d'années[79], la poule se prête bien aux études de conservation d'entités génomiques à travers l'évolution, l'hypothèse étant qu'une telle conservation indique un rôle important pour l'entité en question. Nous avons utilisé ce principe à différentes reprises au cours de la thèse, sur les gènes d'ARN longs non-codants, en étudiant leur conservation chez l'humain et également la souris.

Le génome de la poule est 3 fois plus petit que celui de l'être humain (1.1 milliards de paires de bases – pb – contre 3.2 milliards pb), ce qui permet éventuellement de séquencer 3 fois plus d'individus, ou de réduire les coûts de séquençage de ce même facteur. Le caryotype du génome de la poule montre 38 chromosomes autosomes et de 2 chromosomes sexuels Z et W ($2n$ = 78)[80], contre 22 chromosomes autosomes et de 2 chromosomes sexuels X et Y chez l'humain ($2n$ = 46). Contrairement aux mammifères, les femelles sont hétérogamétiques (ZW) et les mâles homogamétiques (ZZ). Le séquençage du génome de la poule est encore incomplet à ce jour, en raison notamment du contenu en GC de certains chromosomes, qui complique leur séquençage[81]. En effet, les chromosomes 29 et 34 à 38 manquent entièrement, et le séquençage du chromosome 16, qui porte différents gènes liés à l'immunité et aux ARN ribosomaux est encore très imparfait[82].

Dans la présente thèse, nous avons utilisé l'assemblage « *Gallus_gallus-5.0* »[83] (2016) pour tous nos travaux, et notamment pour enrichir l'annotation Ensembl v94, prise comme référence, en gènes d'ARN longs non-codants, comme nous le verrons plus loin. Suite à la sortie de l'assemblage « *GRCg6a* », publié en janvier 2019 sur *Ensembl*, nous avons également proposé cette annotation enrichie en gènes d'ARN longs non-codants en prenant comme annotation de référence l'annotation v100, qui correspondait à ce nouvel assemblage (*cf.* article 1).

II – La régulation de l'expression des gènes : un levier de la variation des caractères complexes et de l'adaptation au milieu

« Their macromolecules are so alike that regulatory mutations may account for their biological difference »

Abstract de King et Wilson en 1975, à propos des humains et des chimpanzés, *in* « Evolution at two levels in humans and chimpanzees », Science 188, 107–116 (1975)[84].

La variation des caractères complexes dans une population est notamment due à des variations d'expression des gènes (Encadré 1). La régulation de cette expression est un processus biologique extrêmement complexe. On sait aujourd'hui que de très nombreux variants sur le génome l'influencent, mais jusqu'à présent, très peu d'entre eux ont été mis en évidence (§ 1). Ces variants régulateurs peuvent agir directement ou indirectement sur l'expression du gène qu'ils contrôlent. On s'intéresse en particulier à ceux qui agissent directement sur l'expression d'un gène, qui sont alors causaux de la variation d'expression et provoquent sous certaines conditions un phénomène appelé « expression allèle-spécifique » (§ 2). L'expression des gènes permet également à l'individu de s'adapter aux variations de son environnement (§ 3).

Encadré 1. L'expression des gènes

La notion d'expression génique, ou expression des gènes, reviendra à de nombreuses reprises dans la présente thèse. Il convient donc d'en avoir une définition claire.

► Au sens le plus large, l'expression des gènes consiste en l'ensemble des processus mis en œuvre pour permettre la transformation de l'information contenue dans la séquence d'ADN en un produit fonctionnel. Dans le cas d'un gène codant pour une protéine, ces mécanismes comprennent notamment : la transcription, les modifications post-transcriptionnelles (épissage, ajout de la queue poly-A, l'export de l'ARN, les mécanismes agissant sur sa stabilité), la traduction, les modifications post-traductionnelles (modification de la protéine par phosphorylation, glycosylation et autres mécanismes agissant sur la stabilité ou le repliement de la protéine). Dans ce sens, la « régulation de l'expression » peut désigner indifféremment la ou les régulations de l'un, plusieurs, ou tous ces phénomènes.

► Dans un sens plus étroit, qui est celui que nous utiliserons dans la présente thèse, l'expression des gènes consiste en la quantité de transcrits (ou un proxy de cette quantité) observable à un instant donné par une technique appropriée (dans notre cas, le RNA-seq) dans un échantillon donné. Cette quantité de transcrits est la résultante de deux phénomènes : la transcription (c'est-à-dire la production de transcrits) et leur dégradation. Dans ce sens, la « régulation de l'expression » désigne simultanément les régulations à l'œuvre dans ces deux grands phénomènes. Notons bien que les techniques d'étude

de l'expression les plus courantes (RNA-seq, microarray, RT-PCRq) permettent d'observer uniquement l'expression des gènes selon ce sens.

► Dans son sens le plus étroit, l'expression des gènes consiste uniquement en le résultat de la transcription d'un locus génique. L'expression en ce sens est mesurable notamment grâce à la méthode GRO-seq (*Global run-on sequencing*), proposée par Core *et al*.[85] (2008), qui permet de connaitre la position, la quantité et l'orientation des RNA polymérases engagées dans la transcription sur l'ensemble du génome.

Il est donc important de prendre garde, en illustrant la régulation de l'expression des gènes par des variants sur le génome et étudiée par RNA-seq, à représenter ces variants comme ayant un effet sur la production de transcrits (transcription) *ou* leur dégradation.

1. Variation de l'expression des gènes et variation des caractères complexes

Le modèle infinitésimal propose que la variation des caractères complexes dans une population est due à une infinité de *loci* ayant de petits effets (§ *a*). On considère aujourd'hui que la majeure partie d'entre eux est localisée dans les régions régulatrices (§ *b*), mais leur nature biologique est inconnue (§ *c*). Cela explique que très peu d'études soient parvenues à en identifier (§ *d*).

On considère pour un individu donné que son phénotype P est influencé par son génotype G, par l'environnement dans lequel il évolue E, et par l'interaction entre cet environnement et son génotype, $G \times E$. Ainsi, on peut écrire :

$$\boxed{P = G + E + G \times E},$$

et la variation des phénotypes $V(P)$ des individus au sein d'une population s'écrit quant à elle

$$\boxed{V(P) = V(G) + V(E) + Cov(G \times E)},$$

où $V(X)$ est la variance de X et $Cov(X, Y)$ la covariance de X et Y.

La composante G est sous-tendue par les effets des différents allèles de *loci* du génome. Ces *loci* peuvent être des *Single Nucleotide Polymorphisms* (SNP), des *INsertion-DELetion* (INDEL), des gènes, etc.

Cette conception tire ses racines du modèle infinitésimal proposé en 1918 par Ronald Fisher[86]. En particulier, Fisher propose avec ce modèle que la composante G des caractères complexes est sous-tendue par une « infinité » de *loci*, ayant chacun un effet « infinitésimal ».

On est donc tenté de se poser quatre questions concernant ces *loci* :

- Combien sont-ils ?
- Où sont-ils ?
- Que sont-ils ?
- Comment agissent-ils ?

a) une variation due à une infinité de loci,

Au début des années 2000, l'ordre de grandeur de l'« infinité » de *loci* sous-tendant la composante G des caractères complexes tournait autour de la dizaine. Par exemple, une étude de 1999 portant sur les allèles impliqués dans l'autisme évoquait « *a large number of loci (perhaps ≥ 15)* »[87]. Le développement des analyses d'association pan-génomiques (*Genome Wide Association Studies*, GWAS) dans le milieu des années 2000 a fait drastiquement augmenter cet ordre de grandeur. Ainsi, en 2014 une méta-analyse d'études travaillant sur la composante génétique de la taille chez l'être humain (utilisant des données sur 253 288 individus) a répertorié 697 variants ayant un effet significatif, expliquant 16% de la variance phénotypique[88].

La faible proportion de variance expliquée par variant était attendue dans le modèle infinitésimal, mais il est étonnant que même pris tous ensemble, les *loci* détectés comme affectant significativement le caractère d'intérêt n'expliquent qu'une proportion très faible de la variation d'un caractère dans la population. Manolio *et al.*[89] (2009), qui travaillaient sur la transmission des maladies parlaient de l'héritabilité manquante. Cette héritabilité manquante a été en grande partie « retrouvée » dans des SNPs pour lesquels l'effet associé était trop faible pour être détecté à cause de la taille trop réduite des dispositifs utilisés alors[90,91]. Ainsi, dans l'étude sur la taille citée plus haut, l'ensemble des SNP dits communs (pour lesquels la fréquence de l'allèle mineur est ≥ 0.01) expliquent 50% de la variance phénotypique[88], et une autre étude portant également sur la taille de 3 925 individus montrait que l'ensemble des SNPs détectés expliquent 45% de la variance[90]. En 2017, Boyle *et al.*[92] ont proposé un modèle « omnigénique », dans lequel ils suggèrent que, eu égard à la forte interconnexion des gènes entre eux, l'ensemble des gènes exprimés dans un groupe cellulaire (par exemple un tissu) impliqué dans un phénotype d'intérêt a un effet sur le phénotype en question. L'idée ici est que même si un gène n'a pas un lien direct avec le phénotype, son action a un effet sur « quelque chose » (l'expression d'un autre gène, la quantité d'une protéine, etc.), et, de proche en proche,

cette action finit par influencer, quoique faiblement (« infinitésimalement ») un gène directement impliqué dans le phénotype d'intérêt.

b) situés dans les régions régulatrices,

En ce qui concerne leur localisation, la vaste majorité des variants associés aux caractères complexes sont situés hors des régions codantes[93,94] : Pickrell[93] estime que seulement 2% à 20% des SNP associés à des maladies humaines sont localisés dans les régions codantes. Plus précisément, Maurano *et al.*[95] (2012) ont observé que 76.6% des SNPs associés par GWAS à des maladies ou des caractères complexes chez l'humain, et localisés hors des régions codantes, sont soit situés dans des régions hypersensibles à la DNase I (57.1%), soit en total déséquilibre de liaison avec un autre SNP dans une de ces régions (19.5%). Ces régions hypersensibles à la DNase I sont des régions ouvertes de la chromatine (donc accessibles à des facteurs extérieurs) et sont connus pour être des régions régulatrices de l'expression[96,97]. On peut donc supposer que ces variants dans les régions non-codantes affectent l'expression des gènes qu'ils régulent, ce qui est maintenant une hypothèse commune dans la littérature : la majorité des variants associés à la variation des caractères complexes sont régulateurs de l'expression. En particulier, le consortium GTEx, qui s'intéresse précisément au lien entre variation génétique et variation d'expression dans les tissus humains[98] à l'aide de 17 832 échantillons issus de 52 tissus et 838 donneurs, a montré que les variants régulant directement l'expression des gènes (dits variants *cis*-régulateurs, voir le point *b)* ci-après) étaient localisés à près de 60% dans des régions introniques, à ~10% dans des régions promotrices, puis en des proportions plus faibles dans d'autres entités génomiques[99]. Enfin, les paires « variant(s) *cis*-régulateur(s) – gène régulé » sont significativement enrichies dans les TAD (*topologically associated domain*)[99], qui sont des segments de l'ADN formant des boucles et au sein desquelles les régions du génome interagissent plus souvent entre elles qu'avec des régions du génome extérieures au TAD[100]. Ces TAD ont chez l'humain une longueur moyenne de 880kb et sont divisés en sous-TAD d'une longueur moyenne de 185 kb[101]. Chez la poule, Foissac *et al.*[102] (2018) rapportent des TAD d'une longueur moyenne de 148 kb dans le foie, et Fishman *et al.*[103] (2019) des TAD d'une longueur moyenne de 308 kb dans des fibroblastes, en utilisant le même algorithme de détection.

c) et de nature inconnue.

L'avant-dernière question que nous posions sur ces *loci*, « que sont-ils ? », est la plus ardue. La nature biologique des variants régulateurs est rarement investiguée, car elle suppose d'identifier

avec précision le *locus* et de valider son implication, ce qui amène à réaliser des manipulations longues et coûteuses. Si on considère que le *locus* consiste en une altération de la séquence d'ADN, on peut spéculer sur le fait que ce soit l'une des différentes micro- ou macro-variations du génome connues (ex. SNP, petits INDEL pour les premiers, CNV pour les seconds[104]). Concernant les SNP, il est important de bien noter que les SNP détectés par GWAS ne sont que des marqueurs, en déséquilibre de liaison avec le variant qui a un effet réel (qui peut être, on l'a vu, un autre SNP, un INDEL ou autre chose), et qu'il est extrêmement rare que le SNP marqueur soit aussi causal.

d) exemple de rares mises en évidence de loci régulateurs de l'expression

La question « que sont-ils ? » se décline aussi en « comment agissent-ils ? ». Y répondre suppose de connaitre la fonction de la séquence affectée par le variant (TFBS, *enhancer*, *silencer*, autre chose ?) et donc de bien connaître l'ensemble des acteurs et mécanismes impliqués dans la régulation de la transcription et de la dégradation des ARN, ce qui est encore loin d'être le cas. En effet, malgré des travaux titanesques comme ceux du consortium ENCODE (*Encyclopedia of DNA Elements*)[105], du *Roadmap Epigenomics Project*[106], ou encore du consortium GTEx[98], les annotations fonctionnelles des différents éléments régulateurs restent très générales, et les mécanismes sous-jacents mal connus[107]. Citons trois exemples de mise en évidence de variants causaux : (*i*) un haplotype à 2 SNP dans une région promotrice régulant l'expression d'un gène codant, (*ii*) un SNP dans une région *enhancer* régulant l'expression d'un ARN long non-codant (classe de gènes sur laquelle nous reviendrons plus longuement dans la deuxième partie de cette introduction), qui d'ailleurs régule lui-même l'expression d'un gène codant, et enfin (*iii*) 2 SNP, l'un dans un promoteur et l'autre dans un intron régulant l'expression d'un gène codant. Pour commencer, citons ainsi la mise en évidence par Le Bihan-Duval *et al.*[108] (2011) d'un haplotype composé de 2 SNP affectant l'expression du gène *BCMO1* et par suite la couleur de la viande du muscle pectoral chez des poulets de chair. Dans cette étude, les auteurs ont commencé par réaliser une analyse « génotype-phénotype » avec pour phénotype d'intérêt la couleur de la viande en utilisant deux populations de poulets de chair divergents, entre autres, pour cette couleur. Ils se sont ensuite intéressés à *BCMO1*, parmi les 30 gènes de la région QTL mise en évidence (4 Mb de longueur), qui était à la fois un bon candidat positionnel (car localisé près du marqueur le plus significatif) et fonctionnel (car impliqué dans la conversion du β-carotène – jaune – en rétinal incolore). Ayant montré que cette région QTL était également une région contrôlant l'expression du gène *BCMO1* lui-même (voir à ce sujet le paragraphe suivant sur les *cis*-eQTL), les auteurs ont

cherché un variant régulateur de l'expression du gène. Ils se sont intéressés aux variants présents dans le promoteur proximal du gène et ont mis en évidence 2 SNP en déséquilibre de liaison complet qui formaient deux haplotypes, l'un fixé dans une lignée (Gn_{57}G, où n_{57} représente 57 nucléotides), et l'autre (An_{57}A) majoritaire dans l'autre lignée. Les auteurs ont alors transfecté des cellules d'une lignée d'hépatocytes de poule (lignée « LMH ») avec un système rapporteur basé sur la luciférase, qui consiste à insérer dans un plasmide la région promotrice de *BCMO1* suivi du gène de la luciférase, qui code pour une enzyme dont l'intensité de fluorescence est une indication du niveau d'expression du gène. En testant les deux haplotypes grâce à ce système, les auteurs ont pu montrer que l'haplotype Gn_{57}G dans le promoteur diminuait significativement l'expression de *BCMO1* par rapport à l'haplotype An_{57}A et aux haplotypes recombinants Gn_{57}A et An_{57}G, qui avaient tous trois le même effet sur l'expression. Il apparait donc clairement que la présence d'un G aux deux positions diminue l'expression du gène, mais les mécanismes sous-jacents n'ont pas été identifiés. Ensuite, chez le chien, Plassais *et al.*[109] (2016) ont pour leur part mis en évidence un SNP régulateur de l'expression d'un ARN long non-codant (*GDNF-AS*), régulant lui-même l'expression d'un gène codant une protéine à proximité (*GDNF*), et connu comme impliqué dans une neuropathie caractérisée par une insensibilité à la douleur. Pour ce faire, les auteurs ont réalisé une GWAS avec 54 animaux (malades et sains) et identifié une région de 1.8 Mb. Ils ont ensuite séquencé cette région chez des individus malades et sains et mis en évidence environ 500 variants. Parmi eux, un seul ségrégeait de la façon attendue dans une population de 300 animaux sains et malades et était absent d'une autre population de 900 chiens sains. La comparaison de la région avec son homologue chez l'humain a montré que le SNP affectait un élément *enhancer*. Les auteurs ont enfin montré que la présence du variant entrainait une forte diminution de l'expression des deux gènes, et que le variant en question altérait fortement la liaison d'éléments régulateurs. Pour terminer avec ces exemples, citons le travail de Tian *et al.*[110](2019). Les auteurs ont identifié chez l'humain 2 SNP régulant l'expression du gène *ATF1*, un oncogène impliqué dans le développement d'un cancer colorectal. En utilisant des données de GWAS déjà existantes dans des populations asiatiques, les auteurs ont identifié 157 gènes codants candidats, répartis à ± 1 Mb de 15 *loci*. En analysant les effets sur la prolifération cellulaire de l'utilisation d'ARN interférents contre ces gènes, les auteurs ont montré qu'*ATF1* avait le plus fort effet sur la prolifération, rôle confirmé par d'autres moyens (sur-expression dans des lignées cellulaires, étude de la prolifération *in vivo* de cellules le sur-exprimant). De plus, *ATF1* était significativement surexprimé dans des jeux de données issus de tumeurs. Les auteurs ont ensuite montré que les deux marqueurs significatifs en GWAS (l'un dans des populations

asiatiques, l'autre dans des populations européennes) étaient également des eQTL d'*ATF1*. En s'intéressant aux SNP en déséquilibre de liaison avec ces marqueurs, les auteurs ont mis en évidence deux SNP, l'un dans le promoteur et l'autre dans le 1er intron, dont les positions superposaient toutes les deux des marques indiquant une accessibilité de la chromatine dans les données ENCODE. Enfin, ils ont montré que l'expression d'*ATF1* variait significativement entre individus selon leur génotype pour ces deux SNP. La prédiction de sites de liaison de facteurs de transcription dans les deux régions concernées a suggéré un rôle potentiel de SP1 et GATA3. La différence d'affinité des deux régions pour ces facteurs de transcription en fonction des allèles des deux SNP a finalement été montrée. Enfin, les auteurs ont montré que *GATA3* et *SP1* agissent de concert et promeuvent l'expression d'*ATF1*, qui lui-même entraîne l'expression de différents autres gènes favorisant la prolifération cellulaire.

On voit à travers ces trois exemples que le repérage des régions contenant les *loci* d'intérêt se fait par GWAS, donc par une approche statistique. Ensuite, si l'étude des fréquences des variants présents dans la région, en lien avec le phénotype d'intérêt, ne suffit pas (cas des exemples 1 et 3), on réduit la région d'étude en s'intéressant aux gènes et à leurs fonctions. Les expérimentations de biologie moléculaire et cellulaire sont un passage obligé pour confirmer le rôle du ou des variants, après avoir confirmé leur effet direct sur l'expression du gène d'intérêt.

2. Les *cis*-régulations et une de leurs conséquences expressionnelles : l'expression allèle-spécifique (ASE)

Un variant régulateur peut agir sur l'expression du gène qu'il régule de différentes manières, et le vocabulaire entourant ces notations n'est pas toujours clair (§ *a*). Ces définitions posées, nous présenterons plus longuement les *trans*- et les *cis*-régulations (§ *b*). Les variants *cis*-régulateurs, auxquels nous nous intéresserons particulièrement dans la présente thèse à cause de leur rôle direct sur l'expression génique, provoquent, lorsqu'ils sont à l'état hétérozygote, un phénomène appelé « expression allèle-spécifique », qu'il est possible d'observer sous certaines conditions (§ *c*).

a) définition des local, cis et trans-régulations

L'expression des gènes est un « phénotype intermédiaire » entre un variant régulateur et un caractère complexe. Prendre ce phénotype, dit « phénotype expressionnel », comme phénotype d'intérêt permet donc de faire un lien entre variant et caractère complexe. Ce faisant, les *loci* mis en évidence comme ayant un impact sur le phénotype expressionnel sont appelés des QTL

d'expression, ou eQTL. Les eQTL sont donc des régions du génome contrôlant l'expression d'un gène, que l'on note alors *eGene*. Comme leur nom le suggère, on peut les détecter par analyse d'association (GWAS), en considérant comme phénotype d'intérêt l'expression d'un gène donné. Les eQTL ainsi détectés sont souvent évoqués dans la littérature avec l'un des préfixes « *cis* », « *trans* » ou « *local* », de manière parfois un peu confuse.

En fait[111–113],

- lorsque le variant détecté par GWAS comme étant associé à la variation d'expression du *eGene* dans la population se trouve à proximité du *eGene* en question, on peut dire que ce variant est un *local*-eQTL.

  On ne peut guère aller plus loin dans la détermination de la nature (*cis* ou *trans*) de la régulation sans manipulation de biologie moléculaire et cellulaire ou analyse de l'expression.

  - Si l'on arrive à montrer que le variant régulateur (qui n'est pas nécessairement le variant analysé par la GWAS) est localisé dans une région régulatrice, et affecte directement l'expression du *eGene*, on pourra alors dire qu'il s'agit d'un *cis*-eQTL.
  - Si non, ou si au contraire on arrive à montrer que le variant régulateur agit indirectement, à travers une molécule intermédiaire, il s'agira d'un *trans*-eQTL.
- En revanche, si le variant est localisé à distance du *eGene*, voire sur un autre chromosome, il est peu probable qu'il agisse directement sur l'expression du *eGene*, et il s'agit là aussi d'un *trans*-eQTL.

b) illustration des trans- et cis-régulations

Illustrons ces définitions à l'aide du cas suivant (voir aussi Figure 3A). Soient V un variant régulateur de nature inconnue (SNP, INDEL, autre chose) en déséquilibre de liaison avec un SNP M (pour marqueur), et G_1 et G_2 deux gènes. Admettons que V, M, G_1 et G_2 sont tous les quatre localisés sur le même chromosome. Admettons également que V contrôle directement l'expression de G_1. C'est donc un *cis*-eQTL du *eGene* G_1. Admettons encore que G_1 contrôle l'expression de G_2 (disons que G_1 est un facteur de transcription). On peut donc dire que V contrôle en *trans* l'expression de G_2 : son action régulatrice est médiée par une molécule intermédiaire, en l'occurrence le facteur de transcription codé par G_1. Pour lever toute

ambiguïté, admettons enfin que dans le cadre d'une GWAS, le variant V ne fait pas partie des marqueurs étudiés, alors que M oui.

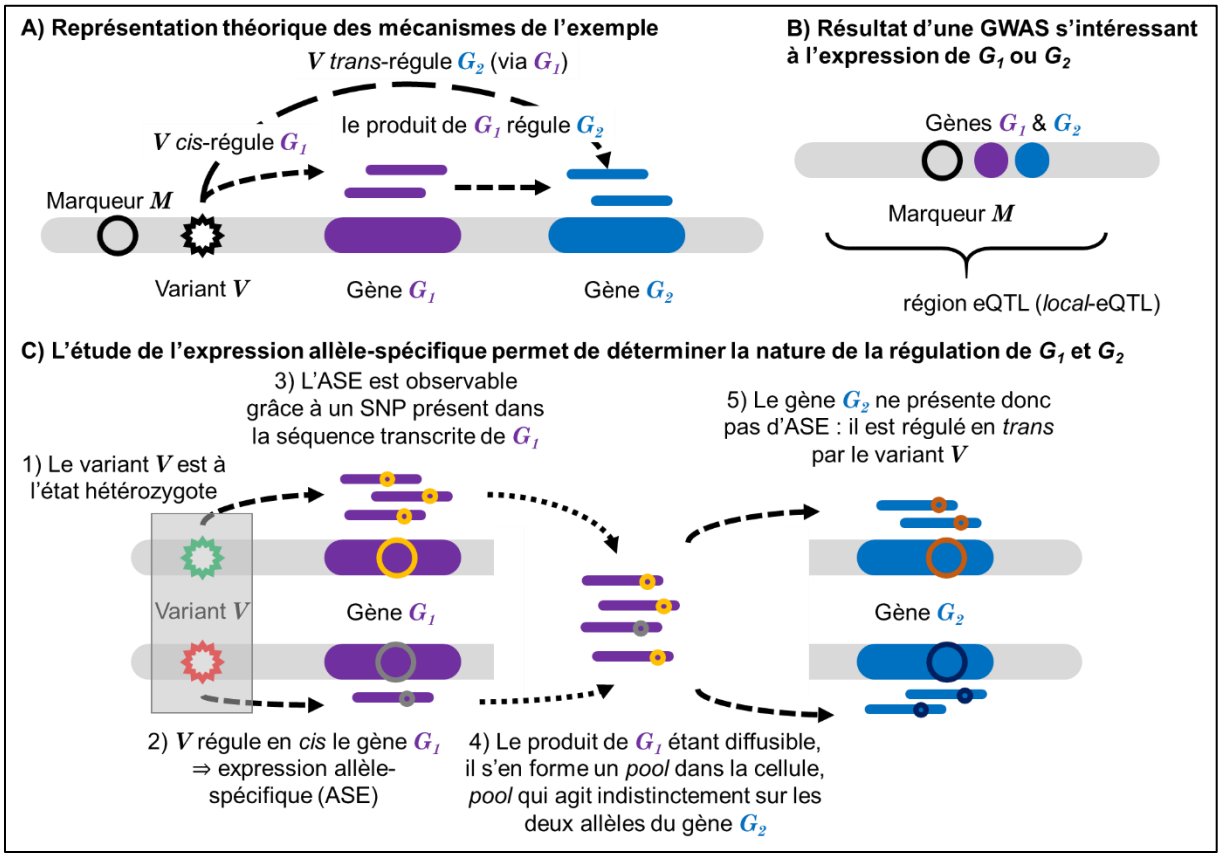


Figure 3 | Illustrations des régulations en *trans*, en *cis* et de l'expression allèle-spécifique, conséquence des *cis*-régulations. A) Illustration de l'exemple proposé dans le texte. Le variant V et le marqueur M, qui a un effet neutre sur l'expression des gènes et est utilisé dans le cadre d'une GWAS, sont en déséquilibre de liaison. V régule en *cis* le gène G_1. Notons que le mécanisme exact n'est pas connu, et que V pourrait très bien se trouver dans une région exprimée de G_1. Le produit de G_1 (un facteur de transcription dans notre exemple) régule ensuite l'expression de G_2. **B)** Le chercheur s'intéressant à l'expression de G_1 ou de G_2 détecterait une région eQTL autour de M et contenant G_1 et G_2. On ne peut que dire à ce stade que c'est une région *local*-eQTL car elle contient G_1, G_2, et, à priori, leurs régions régulatrices. **C)** Pour déterminer la nature de la régulation, il faut étudier l'expression allèle-spécifique des deux gènes dans un contexte où V est à l'état hétérozygote. Notons que la nature et la position de V sont inconnues, d'où le rectangle translucide qui le recouvre. M est omis pour simplifier le schéma. L'observation d'une expression allèle-spécifique pour G_1, détectée grâce à la présence d'un SNP hétérozygote dans la région transcrite de G_1, permet d'affirmer qu'il est *cis*-régulé. Le produit de G_1 étant diffusible, il s'en forme un *pool* qui va réguler l'expression des deux allèles de G_2 de la même façon. Il n'y a donc pas d'ASE pour G_2 (étudiable grâce au SNP hétérozygote dans sa région transcrite), et on peut dire qu'il est *trans*-régulé.

Résumons (Figure 3A) :

$V \rightarrow G_1$: V contrôle directement l'expression de G_1, V est un *cis*-eQTL de G_1.
$V \rightarrow \ldots \rightarrow G_2$: V contrôle indirectement l'expression de G_2, V est un *trans*-eQTL de G_2.

Le chercheur souhaitant étudier la régulation de l'expression de G_1 ou de G_2 et faisant une analyse d'association détectera une région dans laquelle M sera le marqueur le plus associé à l'expression de G_1 ou de G_2, et contenant le variant V (Figure 3B). Cette région pourra être déterminée en prenant les *N* kb amont et les *N* kb aval du SNP le plus significativement associé à la variation d'expression des gènes G_1 et G_2, dans notre cas le SNP M, ou bien, si la densité en marqueurs est élevée, la région sera constituée du continuum des SNP dépassant un certain seuil de significativité. À ce stade, la région contenant V et les deux gènes est un *local*-eQTL. Mais comment déterminer la nature *cis* ou *trans* de la régulation ? Notons d'emblée que si G_2 n'était pas localisé sur le même chromosome que V, la question serait réglée et on pourrait déduire que la régulation se fait en *trans*, mais ce n'est pas le cas ici.

Pour répondre à la question, il faut se rappeler qu'un *cis*-eQTL est situé dans une des régions régulatrices du *cis-eGene* dont il contrôle l'expression, alors que l'action d'un *trans*-eQTL est médiée par une molécule intermédiaire, à priori diffusible (Figure 3C).
Ainsi, si on considère une cellule qui contiendrait le variant V à l'état hétérozygote (avec disons, un allèle augmentant fortement l'expression de G_1 et l'autre allèle sans effet particulier), on devrait observer une différence d'expression entre l'allèle de G_1 porté par un chromosome et l'allèle porté par l'autre chromosome, V étant un *cis*-régulateur de G_1. Au contraire, V étant un *trans*-régulateur de G_2, agissant dans notre exemple à travers le facteur de transcription codé par G_1, ce facteur peut diffuser librement dans la cellule, et affecter de la même manière les deux allèles de G_2. On observera donc la même expression, quel que soit le chromosome. Ainsi, un gène régulé en *cis* présente une différence d'expression entre chromosomes lorsque son variant régulateur est à l'état hétérozygote. On parle d'expression allèle-spécifique, ou ASE (*Allele-Specific Expression*). Notons pour finir que pour « observer » cette différence d'expression entre chromosomes, il faut que les transcrits soient différents selon leur chromosome d'origine : il faut donc qu'ils portent au moins un variant qui soit en déséquilibre de liaison avec le variant régulateur (Figure 3C).

c) l'expression allèle-spécifique (ASE) : une conséquence des cis-régulations

L'ASE consiste donc en une expression différente entre les deux allèles d'un gène.
On sait depuis les années 1960 que l'expression d'un gène peut ne pas être identique selon le chromosome sur lequel il est porté. Le cas le plus évident concerne l'expression mono-allélique de tous les gènes d'un chromosome : le chromosome X. En effet, chez les mammifères femelles,

qui ont un caryotype XX, l'un des deux X est aléatoirement inactivé dans chaque cellule, devenant le corpuscule de Barr[114], et aucun des allèles des gènes portés par ce chromosome X inactivé ne s'exprime. Un autre phénomène assez connu depuis les années 1990[115] est l'empreinte parentale (*imprinting*[116]), qui est une expression différentielle selon l'origine parentale du chromosome, mais pour laquelle le « choix » du chromosome inactivé n'est plus aléatoire, mais déterminé par son origine parentale. Ainsi, dans le cas d'*IGF2* par exemple, seule l'allèle paternel est exprimé[117]. L'*imprinting* semble toucher assez peu de gènes : on en dénombre une quarantaine chez l'humain[118], une centaine chez la souris[119]. En revanche, les chiffres concernant l'expression mono-allélique aléatoire pour les gènes portés par les autosomes sont assez confus. Ce type d'expression concernerait 400 des 3000 gènes testés chez l'humain par Gimelbrant *et al.* en 2007[120] (soit 13%), plus de 10% des 1300 gènes testés chez la souris par Zwemer *et al.*[121] (2012), alors que Reinius *et al.*[122] (2016), en trouvent moins de 1%, toujours chez la souris. Chez la poule en revanche, ce phénomène semble absent[123–126].

Toujours est-il que les exemples précédents sont des cas particuliers d'ASE[127]. En effet, dans le cas le plus courant, l'ASE est moins prononcée, c'est-à-dire que les deux allèles sont exprimés à des niveaux différents. Pour étudier l'ASE, il est fondamental de pouvoir associer à chaque transcrit d'un gène d'intérêt un chromosome d'origine[128]. Pour ce faire, on se fie aux polymorphismes présents dans les transcrits (et à l'état hétérozygotes dans l'individu), en faisant l'hypothèse que ceux-ci sont préférentiellement associés au variant régulateur. Dit autrement, on fait l'hypothèse que variant(s) régulateur(s) et polymorphismes dans les régions exprimées sont en déséquilibre de liaison.

En ce qui concerne la fréquence de l'ASE, et donc des *cis*-régulations, rappelons que l'ASE, qui est l'effet visible d'une *cis*-régulation ne peut être observée que lorsque le variant *cis*-régulateur et au moins un variant dans une région exprimée du gène *cis*-régulé sont à l'état hétérozygote. Les gènes présentant au moins un SNP hétérozygote seront appelés « gènes analysables » dans la suite. Entre différentes études, il nous semble que les variations du nombre de gènes analysables sont plutôt liées à la diversité génétique de la population d'où sont issus les animaux ainsi qu'au nombre d'individus analysés. Ainsi, on s'intéressera dans la suite aux pourcentages de gènes présentant une ASE par rapport au total des gènes analysables, en faisant l'hypothèse que ces gènes analysables constituent un bon échantillonnage de l'ensemble des gènes exprimés. Chez le bovin, Chamberlain *et al.*[129] (2015) ont testé environ 8 000 gènes analysables (3 SNP hétérozygotes par gène en moyenne) dans 18 tissus d'un seul animal, et ont estimé qu'environ 90% de ces gènes (7 067 gènes au total) présentaient une ASE dans au moins un tissu. Ils ont observé une assez grande disparité dans le pourcentage de gènes ASE entre

tissus : selon le tissu, entre 16% et 82% des gènes exprimés et analysables présentaient une ASE (16% dans le thymus, 82% dans le poumon), mais il est difficile de généraliser ces résultats obtenus sur un seul animal. Toujours chez le bovin, Guillocheau *et al.*[130] (2019) ont mis en évidence en utilisant 19 animaux une ASE pour 20% environ des gènes analysables dans le muscle (2 119 sur 12 269) ce qui est cohérent avec les résultats obtenus par Chamberlain *et al.* dans ce tissu (31% de gènes ASE : 1 394 sur 4 455). Chez la poule, Zhuo *et al.* (2017) ont observé dans du cerveau et du foie d'embryon que ~16.3% et 15.4% des gènes analysables présentaient une ASE. Ren *et al.*[131] (2020) ont pour leur part utilisé des échantillons de foie et de muscle pectoral prélevés à trois âges (1 jour, 28 jours et 56 jours) sur des animaux F_1 issus du croisement d'un poulet de chair mâle *Recessive white* et d'une poule pondeuse *Lohmann pink* (n = 6 animaux F_1 pour les âges 1 jour et 28 jours, n = 7 pour l'âge 56 jours). Les auteurs ont considéré comme analysables des gènes portant au moins deux SNP pour lesquels chaque parent était homozygote pour un allèle différent, retenant ainsi 465 gènes. Parmi ces 465 gènes, les auteurs ont observé dans le muscle 2.1%, 0.4% et 1.7% de gènes ASE à 1, à 28 et à 56 jours d'âge, respectivement, et dans le foie, 2.8%, 4.1% et 1.1% de gènes ASE. Enfin, profitant des différents tissus et âges, Ren *et al.* n'ont identifié qu'un seul gène ASE dans les deux tissus (ENSGALG00000037671 à 1 jour d'âge). En revanche, ils ont identifié 5 gènes ASE à deux âges et 4 gènes ASE aux 3 âges. Chez le porc, Maroilley *et al.*[132] (2017) ont observé dans le sang total de 38 animaux une ASE pour environ 11% des gènes analysables, en ne considérant comme ASE que les gènes contenant au moins un SNP hétérozygote présentant une ASE dans au moins un tiers des animaux hétérozygotes. Chez l'humain, Edsgärd *et al.*[133] (2016) ont observé que 13.6% des gènes analysables présentaient une ASE (1 389 gènes ASE sur 10 231 gènes analysables) en utilisant des globules blancs issus de 8 donneurs humains. Toujours chez l'humain, Fan *et al.*[134] (2020) ont détecté 4.7% de gènes ASE (304 gènes ASE sur 6 540 gènes analysables) dans 121 échantillons de reins, et 42% et 57% gènes ASE (2 503 et 2 580 sur 5 961 et 5 465 gènes analysables, respectivement) dans deux populations cellulaires de macrophages. Enfin, le consortium GTEx[99] a observé chez l'humain que 94.7% des gènes codants de protéines et 67.3% des gènes d'ARN longs non-codants étaient *cis*-régulés, en utilisant 52 tissus. Dans le détail, le consortium a observé à travers ces 52 tissus une médiane de 42% des gènes *cis*-régulés, avec un 1er quartile à 26% et un 3^{e} quartile à 57%, mais note une corrélation très claire entre le nombre de gènes détectés comme *cis*-régulés dans un tissu et le nombre d'échantillons disponibles pour le tissu en question (corrélation de Spearman = 0.95).

3. Expression des gènes et adaptation au milieu

Nous avons vu l'importance de la régulation de l'expression dans la variation des caractères complexes dans les populations. Mais l'expression des gènes est également un des leviers de l'adaptation à des conditions changeantes (§ *a*). Nous illustrons ce point par un exemple en présentant l'enchaînement des évènements en réaction à une variation de la concentration cellulaire en cholestérol (§ *b*).

a) généralités

Les cellules, qu'elles constituent par elles-mêmes un organisme unicellulaire ou soient une partie d'un organisme pluricellulaire, sont amenées à faire face à des modifications plus ou moins importantes de leur milieu. Elles doivent s'y adapter pour assurer leur propre survie et, dans le cas des organismes pluricellulaires, celle de l'organisme entier. Lorsque la modification du milieu est détectée ou que l'information arrive à la cellule via les voies de signalisations extracellulaires, des processus biologiques comme le métabolisme, le cycle cellulaire ou l'expression des gènes sont altérés, permettant l'adaptation de la cellule, c'est-à-dire l'ajustement des fonctions biologiques aux nouvelles conditions de vie.

En ce qui concerne l'adaptation des animaux d'élevage aux variations de leur environnement, différents travaux s'y sont intéressés sous l'angle transcriptomique. Parmi les facteurs de stress, citons la température[135,136], les maladies infectieuses[137] ou encore la variation de qualité de l'aliment[138], tous trois en lien plus ou moins étroit avec le changement climatique[4]. Nous avons par ailleurs étudié l'effet du dernier facteur et son lien avec l'efficience alimentaire dans le cadre de l'article 4 de la présente thèse. Même si les variations transcriptomiques ne sont pas les seules impliquées dans la réponse des organismes aux conditions changeantes[139], elles n'en sont pas moins nécessaires à l'ajustement du métabolisme d'une cellule ou d'un organisme aux conditions externes[140]. En effet, si la modification de l'expression génique n'est pas forcément le premier évènement d'adaptation à un stress, comme dans le cas d'un choc thermique où elle est précédée (et par la suite, activée) par l'activation des voies MAPK et PKB/Akt[141], elle est un évènement clef dans l'adaptation, puisque qu'elle seule permet la synthèse de nouvelles protéines.

b) exemple du cholestérol

De manière générale, la perturbation du milieu est détectée par un système idoine, qui induit la transduction d'un ou plusieurs signaux qui modifient *in fine* l'expression de gènes de réponse.

Illustrons ce processus en observant les conséquences d'une perturbation qui entrainerait un diminution de la quantité de cholestérol[142], une molécule hautement régulée. Lorsque la quantité de cholestérol est normale, du cholestérol est lié à la protéine SCAP (*SREBP cleavage activating protein*), elle-même liée à SREBP1 ou SREBP2 (*sterol regulatory element-binding protein*). Ce dernier est considéré comme le facteur de transcription clef de la régulation de la quantité de cholestérol. Ceci permet à SCAP, par un changement de conformation, de se lier avec INSIG (*Insulin-induced gene 1 protein*)[143], elle-même liée à des oxystérols[144]. On est donc en présence d'un trimère SREBP2–SCAP–INSIG localisé dans la membrane du réticulum endoplasmique. Lorsque la quantité de cholestérol diminue (→ perturbation du milieu), il n'y a plus assez de stérols pour se lier à SCAP, provoquant la dissociation de SCAP–SREBP2 et de INSIG (→ détection par le système). SCAP–SREBF2 est alors dirigé vers l'appareil de Golgi[144], où une partie de SREBP2 est clivée par les protéases résidentes[145]. Cette partie est à son tour dirigée vers le noyau (→ transduction du signal) et active l'expression de gènes impliqués dans la synthèse du cholestérol (→ modification de l'expression), c'est bien ici qu'intervient la régulation de l'expression dans l'adaptation aux conditions. L'augmentation de la quantité de ce dernier constitue un rétrocontrôle négatif sur ce processus[146].

III – Annotation fonctionnelle des génomes par RNA-seq et étude de régulateurs de l'expression

Dans cette troisième partie, nous faisons le point sur la méthode phare de la présente thèse, le RNA-seq (séquençage de l'ARN), et exposons, après avoir brièvement décrit son principe, ses applications plus ou moins courantes. Nous verrons que les applications en question reposent toutes sur la double capacité du RNA-seq à accéder au niveau d'expression des transcrits *et* à leur séquence (§ 1.). Nous présenterons ensuite sous divers aspects une classe d'ARN à laquelle nous nous sommes particulièrement intéressés pour son rôle dans la régulation de l'expression des gènes : les ARN longs non-codants (LNC, § 2.).

1. Le RNA-seq permet d'accéder au niveau d'expression et à la séquence des régions exprimées

Le RNA-seq est une méthode permettant d'accéder à la séquence des ARN et à leur niveau d'expression. Comme toute méthode, le RNA-seq a ses particularités et ses outils (§ *a*). Le RNA-seq possède une large gamme d'applications que nous listons, en séparant les applications les plus répandues (§ *b*), que nous avons utilisées pour certaines, des applications plus rares, auxquelles nous nous sommes également intéressés dans la présente thèse (§ *c*).

a) brève description de la méthode

Le RNA-seq (*RNA-sequencing*, séquençage de l'ARN) permet d'accéder aux séquences de l'ensemble des molécules d'ARN (c'est-à-dire les transcrits) présentes dans un échantillon, en proportion de leur quantité dans cet échantillon. La méthode est apparue dans la bibliographie début 2008, avec 3 articles publiés en l'espace de quelques mois, présentant des résultats sur *Arabidopsis thaliana*[147], la levure[148] et les mammifères[149], encore que le premier article parlant d'une analyse de transcriptome par « *High throughput sequencing-by-synthesis* » date de 2006[150]. La méthode, dans ses applications liées à l'analyse de l'expression (voir aussi plus loin « applications courantes du RNA-seq ») est aujourd'hui utilisée en remplacement des puces à ADN (en l'occurrence, à cDNA), appelées aussi *microarray*.

Le RNA-seq consiste donc dans le séquençage d'ARN issus soit d'échantillon de tissus, un mélange hétéroclite de cellules de différents types, soit de cellules individualisées, dans le cas du *single cell* RNA-seq (scRNA-seq) que nous ne développerons pas. En général, ce ne sont

pas l'ensemble des ARN qui sont séquencés : en effet, 95% des ARN d'une cellule sont des ARN ribosomiques (rRNA) ou des ARN de transferts (tRNA)[151]. On peut donc, avant le séquençage, faire une sélection pour ne retenir que les ARN ayant une queue poly-A (pour les enrichir en ARN messagers matures et en certains ARN longs non-codants), faire une sélection sur la taille (pour les enrichir en micro-ARN), ou une ribo-déplétion, qui permet d'éliminer ces rRNA et de retenir ainsi en majorité les ARN messagers et longs non-codants[152]. Les ARN sont ensuite rétro-transcrits en ADN complémentaires (cDNA) – toutes les méthodes « -seq » actuelles consistent à séquencer de l'ADN – et ces cDNA passent par une étape d'amplification par PCR.

En ce qui concerne le séquençage, différentes méthodes existent, et une revue de Goodwin *et al.*[153] (2016) les récapitule. Retenons qu'elles peuvent être divisées en deux grands groupes : les méthodes générant des « *reads* » courts (*short reads*, moins de 1kb, souvent autour de 150 pb actuellement), aujourd'hui les plus répandues et qui correspondent au séquençage de bouts de cDNA, et celles produisant des *reads* longs (*long reads*, de plusieurs kilo-bases), plus récentes que les précédentes, et qui visent à séquencer la molécule entière. Dans le cadre de la présente thèse, nous avons utilisé la méthode *short reads* d'Illumina, le leader sur le marché des appareils de séquençage[154]. Le séquençage permet de transformer l'information biologique portée par la séquence d'ARN en une information numérique sous la forme de *reads*.
Les *reads* sont ensuite alignés sur le génome de référence de l'espèce dont est issu l'échantillon à analyser (sauf dans le cas où il n'y a pas de génome de référence, auquel cas il faut faire un assemblage *de novo* du transcriptome[155–157], que nous ne détaillerons pas ici). L'objectif de l'alignement des *reads* est de déterminer la position de la séquence du *read* sur le génome de référence, et ainsi de savoir de quel gène il est issu. Les premiers outils d'alignement ont été créés pour des données de DNA-seq (séquençage du génome entier), et ne sont pas adaptés à l'alignement de données de RNA-seq (alors que l'inverse est vrai).
La différence fondamentale entre *reads* de DNA-seq et *reads* de RNA-seq est que ces derniers sont issus en grande partie de transcrits matures, c'est-à-dire ayant subi l'épissage. Ainsi, certains *reads* venant de données de RNA-seq chevauchent des jonctions exon-exon, et des nucléotides contigus sur le *read* sont séparés sur le génome par un intron de plusieurs kilobases (en moyenne, 2.1kb pour les oiseaux[158], 3.4kb pour les humains[159]). Les outils d'alignement doivent donc être capables « d'ouvrir » le *read* lorsqu'ils rencontrent un intron afin d'aller chercher la suite plus loin sur le génome. Parmi les différents outils existant pour l'alignement de *reads* de RNA-seq, citons Bowtie2[160], kallisto[161] ou encore STAR[162]. Ce dernier est

actuellement la référence définie par le projet ENCODE[163], et c'est celui qui a été utilisé pour aligner les données de la présente thèse.

b) applications courantes du RNA-seq : étude de l'expression, modélisation de nouveaux gènes et étude de l'editing

On peut distinguer quatre grands types d'applications courantes pour le RNA-seq :

- l'étude de l'expression des gènes, avec différentes déclinaisons,
- la modélisation de nouveaux transcrits et de nouveaux gènes,
- la détection de variants présents dans les séquences transcrites, avec en particulier l'étude de l'*editing*,
- l'étude de l'expression allèle-spécifique, qui nécessite à la fois l'étude de l'expression et la détection de variant

Encadré 2. Normalisation de l'expression des gènes

L'idée de l'évaluation de l'expression des gènes est de déterminer le nombre de transcrits issus de chaque locus génique connu dans un échantillon. Cependant, les comptages bruts des *reads* ayant été alignés sur un gène donné ne fournissent pas une bonne estimation de cette grandeur, ne permettent pas de comparer l'expression de deux gènes dans un même échantillon, et encore moins de comparer l'expression des gènes entre échantillons.

Deux facteurs importants influencent en effet le nombre de *reads* comptés comme alignés sur un gène donné : (1.) la taille du gène et (2.) le nombre total de *reads* qui ont été alignés :

► Plus un gène est grand, plus les transcrits qui en sont issus sont grands, et plus ils « captent » de *reads*. Ainsi, à niveau d'expression équivalent (en termes de nombre de transcrits) un grand gène captera plus de *reads* qu'un petit gène.

► Ensuite, plus le nombre de *reads* générés par le séquenceur pour un échantillon donné est grand, on parle aussi de « taille de la librairie », plus le nombre de *reads* alignés est important et plus il y aura de *reads* alignés sur chaque gène.

Il faut donc, pour pouvoir comparer l'expression des gènes entre des échantillons, normaliser par la taille des gènes et par la taille de chaque librairie.

Deux méthodes très utilisées à ces fins de normalisation sont la méthode RPKM[149] (*Reads Per Kilobase of transcript per Million reads mapped*) et la méthode TPM[164] (*Transcripts Per Million*, qui prend bel et bien en compte la taille du locus malgré ce que son nom peut laisser croire).

Elles se calculent de la manière suivante :

Soit g un gène de longueur L_g kilobases. On considère une librairie de taille N *reads*, répartis sur un total de G gènes, et dans laquelle n_g reads sont associés au gène g.

Soient RPKM_g et TPM_g les expressions en RPKM et en TPM associées au gène g.

On a :

$$\mathrm{RPKM}_g = \frac{n_g}{L_g} \times \frac{1}{N} \times 10^6$$

et :

$$\mathrm{TPM}_g = \frac{n_g}{L_g} \times \frac{1}{\sum_{i=0}^{G} (n_i \div L_i)} \times 10^6$$

On voit donc que le RPKM rapporte, pour un gène donné, le nombre de *reads* normalisé par la taille du locus (terme de gauche) à la taille totale de la librairie (terme du milieu) et multiplie cette grandeur par un million (terme de droite). C'est l'expression la plus simple pour la prise en compte des deux facteurs évoqués plus haut, mais il présente l'inconvénient de varier significativement entre réplicats biologiques[164].

Le TPM rapporte pour sa part, pour un gène donné, le nombre de *reads* normalisé par la taille du locus (terme de gauche) à la somme de cette même grandeur pour tous les gènes (terme du milieu), et multiplie par un million (terme de droite). Ainsi, les TPM expriment la proportion de *reads* normalisés associés à un gène dans l'ensemble de la librairie (en partie par million). La somme des TPM de n'importe quel échantillon est donc égale à 1 million, ce qui rend aisée la comparaison de l'expression d'un gène entre deux échantillons, contrairement aux RPKM puisque les tailles de librairies sont généralement différentes.

Notons tout de même que ces comparaisons ne devraient pas être réalisées entre échantillons dans lesquels la concentration d'ARN ou la proportion des différents transcrits sont différentes[165].

Pour terminer, le package *edgeR*[166], utilisé pour les analyses d'expressions différentielles, applique aux comptages bruts une normalisation appelée TMM (*Trimmed Mean of M values*), reposant sur l'hypothèse que d'une librairie à l'autre, la majorité des gènes devrait avoir le même niveau d'expression. Ainsi, edgeR multiplie chaque librairie par un facteur visant à ramener les comptages bruts de la majorité des gènes au même niveau qu'une librairie prise en référence, en retirant du calcul les gènes les plus différentiellement exprimés entre

librairies, et les gènes les plus exprimés[167]. La normalisation TMM retire donc les 30% de gènes les plus différentiellement exprimés et les 5% de gènes les plus exprimés[168]. Notons l'hypothèse biologique faite par cette méthode : la plupart des gènes ne sont *pas* différentiellement exprimés entre deux échantillons et une minorité de gènes sont très fortement exprimés.

Étude de l'expression : méthodes – L'étude de l'expression consiste à évaluer le niveau d'expression des gènes à partir du nombre de *reads* de RNA-seq alignés sur chaque gène connu. Cette étude se fait presque toujours à l'échelle des gènes, et non de leurs transcrits, car notre connaissance des génomes, notamment en termes de transcrits, est loin d'être parfaite, et elle est particulièrement lacunaire chez les espèces d'élevage. Ainsi, la version 100 d'Ensembl (datant d'avril 2020) compte pour l'être humain 228 116 transcrits pour 44 438 gènes (20 438 codants, 24 000 non-codants dont 16 907 longs), soit 5 transcrits par gène environ, *versus* pour la poule, 39 288 transcrits pour 24 044 gènes (16 878 codants, 7 166 non-codants dont 5 506 longs), soit 1.6 transcrits par gène environ. Tous les transcrits sont donc loin d'être connus. Quand bien même ils le seraient, les algorithmes actuels, basés sur les *reads* courts (*short reads*) ne permettent pas d'associer correctement une expression à un transcrit. Ainsi, certaines méthodes de quantification de l'expression comme featureCount[169] fournissent une évaluation de la quantité de *reads* associés à chaque gène (ou d'ailleurs n'importe quelle entité génique, comme un exon) en comptant directement le nombre de *reads* qui le chevauchent. D'autres méthodes comme RSEM[170], qui est un outil de référence du projet ENCODE, utilisent un modèle statistique pour inférer la quantité de *reads* associée à chaque transcrit connu avant d'agréger ces quantité à l'échelle du gène, ce qui permet de réduire les biais techniques[171]. Quelle que soit la méthode choisie, les résultats de ces évaluations de l'expression des gènes peuvent être fournis dans différentes unités, les plus classiques étant le nombre brut de reads (*raw counts*), les TPM (*Transcripts Per Million*) ou les RPKM (*Reads Per Kilobase of transcript per Million reads mapped*), que nous avons détaillés dans l'Encadré 2.

Étude de l'expression : applications – En pratique, on se sert de ces résultats d'évaluation de l'expression à diverses fins. D'abord, pour qualifier le niveau d'expression d'un gène ou fixer la limite entre le bruit et l'expression réelle. On considère par exemple un gène codant une protéine comme exprimé à partir de 1 TPM environ, seuil qui descend, comme on le verra plus loin, à 0.1 TPM pour un gène d'ARN long non-codant.

Ensuite, on peut se servir des comptages des *reads* afin de comparer l'expression des gènes entre des échantillons issus de conditions différentes (par exemple, groupe « nourri avec un aliment hypo-énergétique » *versus* « contrôle » comme dans l'article 4) : il s'agit de l'analyse de différentiel d'expression, ou analyse DE. C'est l'application classique du RNA-seq. L'idée est de détecter les gènes significativement différentiellement exprimés entre les groupes, en faisant l'hypothèse que tout ou partie de ces gènes jouent un rôle dans, ou sont affectés par, ce qui différencie les groupes. Là encore, différents outils existent pour réaliser ce type d'analyse, les plus connus étant DESeq2[172] ou edgeR[166]. C'est ce dernier que nous avons utilisé dans la présente thèse. Cet outil normalise les comptages bruts par la méthode TMM (*Trimmed Mean of M-values*, voir Encadré 2) avant de réaliser l'analyse DE proprement dite. Certains de ces outils ont d'abord été conçus pour être utilisés avec des données de microarray, qui permettaient également de faire des mesures d'expression. C'est en revanche la seule application possible du microarray sur les quatre que nous détaillons ici.

Enfin, on peut également étudier les corrélations d'expression entre gènes, en faisant l'hypothèse que des gènes fortement corrélés en expression (« co-exprimés ») partagent un lien biologique. On peut donc par exemple inférer un rôle pour un ou des gènes encore peu connus, co-exprimés avec des gènes mieux connus (voir par exemple l'article 1). Des méthodes existent pour étudier les corrélations deux à deux entre les gènes connus exprimés d'un tissu (ordre de grandeur de 10 000 à 15 000 gènes), puis sélectionner et regrouper des gènes partageant de fortes corrélations. Citons particulièrement la méthode WGCNA[173] (*Weighted Gene Co-expression Network Analysis*). Cette méthode élève les corrélations à une puissance (souvent autour de 6), permettant de « creuser » la matrice de corrélation (seules les corrélations proches de 1 restent relativement élevées, les autres sont fortement diminuées). Cela permet de dégager des groupes, dits « modules », de gènes fortement co-exprimés. Nous avons utilisé cette méthode à différentes reprises dans cette thèse (voir articles 3 et 4).

Modélisation de nouveaux transcrits et gènes – Pour étudier un gène, il faut d'abord en connaître l'existence. Le RNA-seq est une méthode de choix pour la modélisation de nouveaux transcrits, qui, superposés, définiront l'étendu d'un gène nouveau ou préciseront les limites d'un gène déjà connu. En effet, le séquençage se fait sans à priori, c'est-à-dire qu'aucune amorce avec une séquence précise n'est nécessaire. On peut ainsi supposer que tous les transcrits d'un échantillon sont séquencés. L'observation visuelle de l'alignement des *reads* résultant du séquençage (à l'aide du logiciel IGV[174] par exemple) montre des empilements (« *pile-ups* ») de *reads* au niveau des exons exprimés des gènes connus, mais également des empilements qui

dépassent les limites en 5' ou en 3' d'un gène, voire des empilements dans des régions intergéniques. Ces empilements de *reads* dans des régions intergéniques peuvent résulter d'un bruit de fond transcriptomique (si un tel évènement existe en effet, et n'est pas lié à des erreurs d'alignement[175,176]). Mais ils peuvent aussi indiquer que la région génomique subit bel et bien la transcription voire l'épissage (dans le cas de *reads* chevauchant des jonctions), et constitue donc un gène jusqu'alors non-modélisé. La modélisation de transcrits (ou assemblage du transcriptome) à partir de données de RNA-seq se heurte à une difficulté similaire à l'évaluation de leur expression : pour un *read* donné, comment savoir à combien, et le cas échéant, à quelle(s), isoforme(s) il appartient ? La solution, comme précédemment, passe par l'utilisation de modèles statistiques pour la modélisation des transcrits avec leurs exons. Pour le gène, il est défini comme la région couverte par l'ensemble des transcrits dont au moins un exon se chevauche sur le même brin. Les transcrits et gènes que nous avons utilisés dans le cadre de l'article 1 avaient été modélisés préalablement[177] à l'aide de l'outil Cufflinks[178], qui fait également partie des outils de référence du projet ENCODE.

Détection de différences entre ARN et ADN – L'édition de l'ARN (*RNA editing*) représente des modifications post-transcriptionnelles de la séquence de l'ARN. Les deux modifications les plus connues sont la transformation d'une adénine en inosine (*A-to-I*), qui est lue comme une guanosine lors de la traduction[179], et la transformation d'une cytosine en uracile (*C-to-U*), même si d'autres modifications ont été observées (*C-to-G*, *G-to-A* par exemple)[180]. La détection d'un phénomène d'*editing* consiste à détecter un variant sur la molécule d'ARN alors que l'ADN génomique est homozygote à cette position. L'*editing* a surtout lieu sur les ARN double-brins, notamment ceux issus des éléments *Alu* chez l'humain[181] (dont les équivalents chez la poule sont les séquences CR1[182]). Les évènements d'*editing* ont tendance à y apparaitre en *clusters* (phénomène dit d'hyper-édition), qui provoquent l'apparition de ≥ 20 sites édités[183]. Au moment de l'alignement des *reads*, l'application de filtres classiques quant aux « *mismatches* » (différences entre séquence du *read* et génome de référence) pour les *reads* de RNA-seq fait que ces *reads* sont éliminés. L'étude de l'*editing* par RNA-seq implique donc d'étudier les *reads* non alignés. Dans une publication de 2017, le consortium GTEx a cherché à bâtir un atlas de l'*editing A-to-I*, à partir de la base de données d'évènements d'*editing* RADAR[184] et de données RNA-seq du consortium (RNA-seq sur 53 tissus de 552 donneurs, pour un total de 8 551 échantillons). Ce travail a produit un catalogue de 2 802 751 sites édités chez l'humain. Cependant, l'étude par tissu a montré qu'au sein d'un tissu donné, une faible proportion de ces sites connus subit un évènement d'*editing* (≤ 0.15% de ces presque 3 millions de sites

répertoriés, soit moins de 4 500 évènements)[185], dont plus de la moitié ont lieu dans des sites hyper-édités. L'ensemble de ces études explique que l'évaluation du nombre d'évènements d'*editing* peut varier de quelques dizaines à centaines[186,187 180,188] lorsque les filtres classiques d'alignement de données RNA-seq sont utilisés (rendant cet évènement rare), à quelques milliers[189], voire quelques millions[185].

c) applications plus rares du RNA-seq : détection de variants génomiques et expression allèle-spécifique

Détection de variants génomiques – À l'*editing* près, le RNA-seq devrait permettre d'accéder aux variants situés dans les régions transcrites du génome, un sous-ensemble des variants détectés par les méthodes de séquençage du génome entier comme le DNA-seq. Cependant, comme nous le verrons dans l'article 2, la littérature à ce sujet est assez rare. En effet, depuis le premier article de Piskol *et al.*[190], publié en 2013, moins d'une dizaine d'études[191–198] se sont attelées à proposer des méthodes ou des outils de détection des SNP par RNA-seq chez les mammifères, et surtout à étudier la concordance entre les SNP détectés par RNA-seq et ceux détectés par DNA-seq (qui est actuellement la référence), ou parfois par puce à SNP[192]. De plus, ces études avaient en général un nombre restreint d'individus (en général ≤ 10[193,195,196]) et ne disposaient pas de données de RNA-seq et de DNA-seq collectées strictement sur les mêmes échantillons biologiques, ni même parfois sur les mêmes individus[191,198]. Nous présenterons dans l'article 2 les résultats que nous avons obtenus dans la détection de SNP par RNA-seq utilisant les méthodes recommandées par ENCODE, et particulièrement la comparaison de ces SNP détectés par RNA-seq à des SNP détectés par DNA-seq sur les mêmes tissus des mêmes individus, issus de deux populations de 15 et 8 poules.

Expression allèle-spécifique – Enfin, s'il est possible de faire des comparaisons d'expression d'un gène entre deux individus ou groupes d'individus, il est également possible de comparer chez un individu diploïde, l'expression d'un gène entre les 2 chromosomes le portant. Lorsqu'une différence d'expression entre les deux allèles est détectée, on parle d'expression allèle-spécifique (*allele-specific expression,* ASE), phénomène décrit plus en détail plus haut. Pour associer à chaque *read* son chromosome d'origine, il faut disposer de la liste des variants hétérozygotes présents dans les régions exprimées du génome et de ceux présents sur les *reads* de RNA-seq. On peut pour ce faire n'utiliser que des données de RNA-seq, car, comme nous l'avons vu plus haut, l'*editing* ne concerne qu'un nombre très faible de sites et ne devrait introduire que de très rares erreurs dans cette détection. Une fois les variants détectés, ce qui

nécessite, rappelons-le, un premier alignement des *reads* de RNA-seq sur le génome de référence et l'utilisation d'outils de détection, il faut aligner à nouveau les *reads* de RNA-seq, mais cette fois-ci sur un génome de référence dit « masqué ». Sur ce génome masqué, les positions des variants détectés sont comme dissimulées à l'aligneur (en pratique, les nucléotides du génome de référence sont remplacés par un *N* qui n'est pas pris en compte dans l'alignement) et n'influencent donc pas l'alignement. L'idée est de ne pas favoriser l'alignement de *reads* portant le même allèle que le génome de référence à des *reads* portant un variant (qui ne serait pas alignés, ou alignés ailleurs), risquant de créer artificiellement un déséquilibre dans le nombre de *reads* associés à chaque chromosome, et donc une fausse « ASE ». Une fois les *reads* alignés, l'évaluation de l'expression allèle-spécifique se fait, en première approche, en comptant à chaque position de SNP hétérozygote, le nombre de *reads* portant chaque allèle. C'est l'approche adoptée par certains articles étudiant l'ASE[129,130,132,199,200]. L'inconvénient de cette approche, utilisée par l'outil ASEReadCounter[201] par exemple, est qu'elle ne permet pas d'obtenir facilement une évaluation de l'expression allèle-spécifique à l'échelle d'un gène, mais seulement à l'échelle de chaque SNP qu'il contient. Une autre approche, adoptée notamment par phASER[202] et ses déclinaisons[203] utilisés par le consortium GTEx (notamment dans le cadre de l'article 5, en préparation), consiste d'abord à phaser autant que possible les SNP portés par un gène. Les comptages sont ensuite faits à l'échelle de ces haplotypes, puis un haplotype est associé à chaque gène (voir aussi le paragraphe 3 de la section II Articles et travaux complémentaires). Tout ce processus permet notamment de réduire le nombre de faux-positifs[202]. Différentes études ont également adopté cette approche consistant à phaser les variants avant de réaliser les comptages, avec phASER ou d'autres outils[204–208].

2. Les ARN longs non-codants (LNC) : des régulateurs encore méconnus

Les ARN longs non-codants (LNC) forment une classe d'ARN dont l'étendue a été découverte relativement récemment grâce aux données de RNA-seq. Nous commencerons par définir et présenter ces gènes sous divers aspects (§ *a*), puis consacrerons un paragraphe à leur expression (§ *b*). Nous présenterons ensuite l'approche que nous avons choisie pour les classifier, (§ *c*) et nous listerons finalement grâce à des exemples les modes d'action connus à travers lesquels ils régulent cette expression (§ *d*).

a) définition et prédiction

Définition – L'histoire de la découverte de l'ARN est complexe, tant sur le plan moléculaire (identification et caractérisation de la molécule du point de vue biochimique) que sur le plan des idées (compréhension de son ou ses rôle(s), « dogme central » de Crick[209,*] en 1957).

La découverte des ARN messagers peut être « datée » de 1961, avec l'article de Brenner *et al.*[213]. Ils ont comme caractéristique majeure (centrale, si l'on ose dire) d'être codants pour des protéines, c'est-à-dire qu'une partie de leur séquence nucléotidique correspond à une séquence d'acides aminés, et que le passage de la première à la seconde se fait via la traduction par les ribosomes.

Les gènes à ARN longs non-codants et leurs transcrits (*long non-coding RNA*, *lncRNA*, et dans la suite, LNC) ne possèdent pas de cadre de lecture ouvert (*Open Reading Frames*, ORF) comparable à ceux des gènes codants des protéines (dans la suite, PCG), ce qui, comme nous le verrons plus loin, permet de les en distinguer. En ce qui concerne leur « longueur », ils ont été définis comme étant des transcrits de taille ≥ 200 nt, ce qui les distingue des ARN courts non-codants. Cette valeur arbitraire pourrait bien venir de ce que les protocoles de purification d'ARN classiques, particulièrement le RNeasy de Qiagen utilisé par Kapranov *et al.*[214] en 2007 ne retiennent pas les ARN de moins de 200 nt[215].

[*] Ce « dogme central » (sous-entendu, de la biologie moléculaire) a une histoire tourmentée. D'abord le terme « dogme » (« Opinion émise comme une vérité indiscutable » pour le dictionnaire Le Robert[210]) a été incorrectement utilisé par Crick qui n'en saisissait pas le sens, comme il l'explique dans son autobiographie : « *I used the word the way I myself thought about it, not as most of the world does, and simply applied it to a grand hypothesis that, however plausible, had little direct experimental support.* »[211]. Ensuite, l'hypothèse de Crick est que l'information, c'est-à-dire la détermination de la séquence, va de l'ADN vers les protéines, sans retour en arrière. Elle a été incorrectement expliquée par Watson en 1965, et on la comprend et l'explique maintenant (toujours incorrectement) sous la forme « l'ADN fait de l'ARN qui fait des protéines »[212].

Un LNC est donc, selon la définition classique retrouvée au début de nombreux articles :

« Un ARN d'une taille ≥ 200 nt et ne codant pas pour une protéine »

Comme nous le verrons dans le paragraphe ***f)***, la majorité des études réalisées sur les LNC ont montré que ces gènes ont des rôles de régulateurs de l'expression d'autres gènes[216–219] (citons pour les LNC les plus connus *XIST*[220], *H19*[221] ou encore *HOTAIR*[222]).
Nous avons par ailleurs travaillé sur des données de RNA-seq obtenues après des protocoles ne retenant que les ARN poly-adénylés (poly-A^+), comme dans la majorité des études concernant les LNC[223]. En effet, beaucoup de LNC sont détectés comme étant poly-A^+, mais il est difficile de trouver des proportions dans la bibliographie, car il semble que, contrairement aux PCG, les LNC tendent à être « bimorphes », c'est-à-dire présents aussi bien à l'état poly-A^+ que poly-A^-[224–227]. Ils semblent par ailleurs être 5'-cappés[218,228]. On peut donc dire que nous nous sommes intéressés dans la présente thèse aux LNC en tant qu'ARN *poly-adénylé, au rôle potentiellement régulateur de l'expression d'autres gènes*, d'une taille ≥ 200 nt et sans ORF comparable à ceux des PCG.

Prédiction – Comme nous l'avons vu plus haut, les LNC ne possèdent pas d'ORF comparables à ceux des PCG. C'est justement en prédisant le potentiel codant d'un transcrit à l'aide d'outils idoines et en le comparant à celui attendu pour un PCG que l'on peut le classer comme étant un transcrit de PCG ou de LNC. Les outils de prédiction des LNC (revus brièvement ici : [177]) utilisent en général le critère « longueur de l'ORF » (la suite de codons entre un codon start et un codon stop), et parfois l'existence dans une base de données de la protéine potentiellement associée. FEELnc[229], utilisé lors d'une précédente thèse[177] pour la prédiction des LNC utilisés dans l'article 1, présente différents avantages par rapport aux autres outils existant. FEELnc utilise pour la prédiction du potentiel codant la fréquence des *k*-mers, en plus de différents paramètres liés aux ORF. Par ailleurs, FEELnc réalise la classification des LNC par rapport au PCG le plus proche, en utilisant la nomenclature de GENCODE, puisqu'il a été développé par Thomas Derrien, auteur de l'article de référence sur les LNC humains[226] en 2012. FEELnc prédit le potentiel codant de chaque transcrit à l'aide d'un algorithme de forêt aléatoire (*random forest*) entrainé à l'aide d'un jeu de PCG et de LNC connus fourni par l'utilisateur, en prenant donc en compte pour l'entrainement et les prédictions de potentiels codants les *k*-mers et les ORF. En l'absence de LNC dans le jeu d'entraînement, il est possible de les remplacer par des LNC « virtuels » créés soit par mélange aléatoire des séquences des PCG, soit par des séquences intergéniques choisies au hasard. Notons pour finir que les critères autour des ORF comparent

les ORF des LNC à celles des PCG connus. Il se trouve en effet que les LNC ont bel et bien des ORF, et certains LNC s'avèrent être à l'origine de la traduction de petits polypeptides (appelés parfois « micro-protéines »), d'une taille médiane évaluée à 43 acides aminés[230], versus une médiane de 361 acides-aminés pour les protéines connues jusqu'à présent[231].

b) structure, origine biologique et conservation

Structure – Les LNC tels que définis plus haut, comme les ARN messagers (ARNm), sont transcrits par l'ARN polymérase II, subissent une 3'-polyadénylation et ont une coiffe en 5'[218,232]. Après leur transcription, les LNC subissent l'épissage, toujours comme les ARNm, raboutant ainsi leurs exons. Les LNC semblent avoir tendance à être composés de deux exons, aussi bien chez l'humain[226,233] que la souris[234], ou encore chez quatre espèces d'élevage (poule, porc, caprin et bovin)[102]. Cependant, cette tendance est à relativiser. En effet, étant moins exprimés que les PCG (*cf.* plus loin), les LNC sont également plus difficiles à modéliser correctement avec des *reads* courts (*short reads*). Ainsi, Lagarde *et al.*[235] ont montré en 2016 que 60% des LNC qu'ils ont analysés chez l'humain n'étaient pas entièrement modélisés en 3' ou 5' à l'aide d'une méthode alliant RACE (*rapid amplification of cDNA ends*) et *reads* longs, le RACE-seq. Les transcrits qu'ils sont parvenus à modéliser par cette méthode avaient un nombre d'exons comparable à celui des PCG, ce qui suggère que cette différence n'existe pas.

Origine biologique – La recherche de l'origine biologique des LNC définis au « sens large » (≥ 200 nt, ne codant pas de protéines) pourrait nous amener il y a environ 4 milliards d'années, aux origines de la vie elle-même. En effet, en 1962, Alexander Rich a fait l'hypothèse que l'ARN pourrait à la fois porter une information et avoir une activité catalytique[236,237]. Cette hypothèse a ensuite été surnommée l'hypothèse du monde à ARN (« *RNA world* ») en 1986 par Walter Gilbert[238]. Elle est soutenue par différentes découvertes datant des années 1980 qui suggéraient que les ARN peuvent avoir une activité enzymatique (en plus de porter une information). Il a été montré chez *Escherichia coli* et *Bacillus subtilis* que l'activité catalytique de la Ribonucléase P était portée par la partie ribonucléotidique de l'enzyme[239], ou encore que chez *Tetrahymena thermopila*, l'ARN ribosomal est capable de s'auto-épisser[240]. Chez les bactéries, des ARN appelés « *riboswitches* » et capables notamment de lier des métabolites pourraient être des « descendants » de systèmes de régulation basés sur l'ARN[241]. Ce sont toutes les observations montrant une activité catalytique des ARN[242], et en particulier une collaboration entre ARN différents[243], qui tendent à supporter l'hypothèse du monde à ARN. Terminons sur le monde à ARN pour signaler que l'hypothèse n'est pas sans soulever des

questions quant à la stabilité de la molécule d'ARN[244] et que d'autres hypothèses existent, par exemple que les ARN aient eux-mêmes été précédés par des molécules d'une autre nature biochimique, aujourd'hui inconnues[245].

Sans aller aussi loin dans le passé, et concernant maintenant l'apparition d'un LNC dans un génome, il semble qu'elle soit principalement due à la perte de fonction d'un gène auparavant codant[177,246,247], aussi appelé « pseudo-gène ». Cette perte peut être due à l'apparition d'une ou plusieurs mutations[246], de réarrangements chromosomiques ou encore à l'apparition d'un élément transposable dans la séquence codante[248,249]. La nuance entre LNC et pseudo-gène parait être dans certains cas purement sémantique : il existe en effet des pseudo-gènes qui sont exprimés et régulateurs de l'expression de PCG[250,251]. Cela étant, il existe également des régions du génome dont la séquence semble être paralogue de gènes fonctionnels, mais qui ne subit jamais de transcription et n'est plus sous l'influence que de la dérive génétique[252], caractérisant alors un « vrai » pseudo-gène, dépouille d'un gène autrefois codant et condamnée à disparaitre (à moins d'un gain de fonction éventuel). De façon générale, l'hypothèse que des LNC sont issus d'un « regain de fonction » d'un pseudo-gène qui était avant un PCG est cohérente avec l'observation que les LNC partagent avec les PCG différentes caractéristiques structurelles (structure exon-intron, queue poly-A, coiffe 5', etc.). L'apparition (ou la disparition) de nouveaux LNC semble être un évènement assez courant : Kutter *et al.*[253] (2012) ont par exemple observé qu'entre le rat et deux espèces de souris (séparés évolutivement par « seulement » 13 à 19 millions d'années pour le rat et les souris, 1 million d'années pour les deux espèces de souris), seuls 28% des LNC intergéniques orthologues entre ces trois espèces (ne chevauchant aucun PCG) voient leur expression conservée dans le foie de ces animaux, contre 55% des gènes codant des protéines. Ce résultat pourrait éventuellement poser la question de leurs rôles potentiels dans le phénomène de spéciation.

Conservation – Les LNC sont en général bien moins conservés en termes de séquence primaire entre les espèces que les PCG[254,255]. Ils subissent néanmoins une pression de sélection plus importante que les introns ou les régions intergéniques, aussi bien au niveau de la séquence du gène même qu'au niveau des sites d'épissages et de la séquence promotrice[254]. Ainsi, les LNC sont composés de « patches » de séquences conservées, entourées de zones apparemment non-conservées[256,257]. Cette notion de « patches » est à rapprocher de celle de petits motifs de nucléotides appelés *k-mers*, que l'on utilise dans le contexte de la génomique pour désigner des motifs de séquences de longueurs k nucléotides (où k est un entier naturel). Ainsi, Kirk *et al.*[258]

ont proposé dans la revue *Nature Genetics* en 2018 un outil analysant la séquence des LNC en se basant sur des *k-mers*, et l'ont utilisé chez l'humain et la souris avec $k = 6$. Ils montrent notamment (*i*) que les LNC peuvent être partitionnés en communautés en se basant sur leur contenu en *k-mer,* (*ii*) que le contenu en *k-mers* est corrélé avec la localisation cellulaire et la liaison aux protéines et (*iii*) qu'il existe des communautés de LNC (réunis en se basant sur leur contenu en *k-mer*) qui peuvent se retrouver à travers 10 espèces de vertébrés, ainsi que chez l'oursin.

On comprend donc qu'il est compliqué d'étudier la conservation des LNC entre espèces par la séquence. Ainsi, la conservation est classiquement étudiée par synténie[257,259–261], c'est-à-dire que l'on utilise les gènes codants situés en 5' et en 3' du LNC d'intérêt comme « bornes » et que l'on observe dans les autres espèces si un LNC se trouve entre les orthologues de ces gènes codants (sans autres gènes codants entre). Si c'est le cas et que les trois gènes sont dans la même orientation les uns par rapport aux autres, on peut suspecter une conservation. Nous avons utilisé cette méthode dans l'article 1 en complément de la conservation de l'expression à laquelle nous reviendrons un peu plus loin. Ainsi, Ulitsky *et al.*[262] (2011) sont parvenus à « réparer » chez le poisson zèbre des phénotypes anormaux qu'ils avaient obtenus en altérant notamment l'épissage de LNC de poisson zèbre, en utilisant les orthologues humains et murins de ces LNC, alors même que les LNC en question partageaient de faibles similarités de séquence.

Enfin, les séquences en amont des *transcription starting sites* (TSS) sont plus conservées entre espèces pour les LNC que pour les PCG[263] (précisément, le pourcentage de séquences de 20 nucléotides alignés entre génomes est plus élevé), et mieux conservés que les gènes eux-mêmes[264] (mesure de conservation par différentes méthodes). On trouve également des paires PCG:promoteur de LNC qui semblent conservées entre espèces[265]. Cela pourrait expliquer qu'un autre aspect de la conservation des LNC est la conservation expressionnelle : ils présentent une forte conservation de leur tissu-spécificité à travers les espèces[266,267], avec d'ailleurs un très grand nombre de LNC spécifiques des testicules.

Ainsi, il nous parait, suite à ce rapide exposé sur la conservation de LNC, que ce sont (*i*) de courts patches de nucléotides et (*ii*) les séquences en amont des LNC qui tendent à être conservés à travers l'évolution plutôt que la séquence primaire, et qui pourraient donc être importants pour l'action des LNC.

c) expression des LNC et tissu spécificité

Le niveau d'expression des LNC est généralement plus faible que celle des ARN messagers (de l'ordre de 10 à 100 fois moins selon les études[226,268–271], voir aussi l'article 1 à ce sujet). Ainsi, alors que pour les PCG, un niveau d'expression minimum souvent utilisé pour considérer un gène comme exprimé est 1 TPM ou RPKM, ce niveau est de 0.1 TPM ou RPKM[226,272–274] pour les LNC, dans un certain pourcentage des échantillons étudiés pour s'assurer de la reproductibilité de l'expression. Certains auteurs ajoutent parfois une contrainte d'un nombre minimum de *reads* couvrant le gène (par exemple 6 chez de Goede *et al.*[273] [2019], 5 chez nous dans l'article 1), ce qui permet d'éviter de considérer comme exprimé des gènes de petite taille (en particulier < 1kb, seuil auquel le fait de diviser la taille du gène en kb dans les formules des TPM et RPKM [*cf.* Encadré 2] a un effet multiplicatif), et peut-être aussi d'ajouter un aspect plus concret aux formules d'expression.

L'expression des LNC est également plus tissu-spécifique que celle des PCG. Cela signifie que les LNC ont une tendance plus forte à être exprimés à un niveau donné dans quelques tissus et à des niveaux bien plus faibles dans les autres, voire exprimés uniquement dans un tissu et pas du tout dans les autres. Les PCG au contraire ont plus tendance à être ubiquistes, c'est-à-dire exprimés à des niveaux relativement proches dans de nombreux tissus (même s'il existe des PCG tissus-spécifiques). Différentes méthodes existent pour évaluer la tissu-spécificité de l'expression des gènes, revues par Kryuchkova et Robinson[275] (2016). Parmi elles, nous avons utilisé le score τ (pour l'article 1), proposé par Yanai *et al.*[276] (2005). Le τ compare l'expression d'un LNC dans chaque tissu à celle du tissu où elle est la plus élevée et s'exprime sur une échelle allant de 0 (gène ubiquiste, exprimé au même niveau dans tous les tissus étudiés) à 1 (gène « tissu-exclusif », exprimé dans un seul et unique tissu). C'est une mesure de tissu-spécificité qui a une bonne corrélation entre jeux de données différents et est robuste aux méthodes de normalisation[275]. Cela étant, pour un gène donné, le score de tissu-spécificité varie très probablement en fonction du nombre et de la nature des tissus étudiés. En revanche, pour des classes de gènes prises dans leur ensemble (PCG, LNC), on s'attend à retrouver des tendances similaires entre projets utilisant des tissus variés, en nombre suffisant. En ce qui concerne les ordres de grandeur de cette tissu-spécificité, Ransohoff *et al.*[277] (2017), en utilisant le score τ avec 30 tissus chez l'humain observent en effet que les LNC sont plus tissu-spécifiques que les PCG. En fixant le seuil de tissu-spécificité à $\tau \geq 0.88$ (on trouve aussi la valeur $\tau \geq 0.90$[278] ou 0.95[270]), 61% des LNC étaient tissu-spécifiques, versus 30% des PCG. Cabili *et al.*[271] (2011), en étudiant 24 tissus humains, ont estimé que 78% des LNC intergéniques (*cf.* paragraphe suivant pour les classifications des LNC) étaient tissus-

spécifiques, contre 19% des PCG en utilisant une métrique basée sur l'entropie qui compare l'expression d'un LNC à travers les tissus à un modèle dans lequel le LNC ne serait exprimé que dans un seul tissu. Les auteurs montrent par ailleurs que cette tissu-spécificité ne résulte pas du plus faible niveau d'expression des LNC, et qu'à niveaux d'expression comparables, les LNC restent plus tissu-spécifiques que les PCG. Le testicule est un organe dans lequel semble se trouver un grand nombre de LNC tissus-spécifiques : 33% des LNC tissu-spécifiques étudiés par Ransohoff *et al*. (2017) avaient leur plus forte expression dans le testicule ou le cerveau, une proportion similaire à celle obtenue par Cabili *et al*. (2011). De même, dans le cadre des travaux du consortium GTEx, Melé *et al*.[272] (2015) ont observé en utilisant 43 tissus humains que moins de 200 gènes étaient exprimés uniquement dans un tissu, dont 95% l'étaient dans le testicule, avec parmi eux 90% de LNC. Chez les espèces d'élevage, très peu de travaux se sont intéressés à la tissu-spécificité d'une façon globale, ne permettant pas de tirer des ordres de grandeurs du nombre et de la proportion de LNC et PCG tissus-spécifiques dans suffisamment de tissus. Chez le porc, en utilisant 9 tissus et la même métrique que Cabili *et al*., Tang *et al*.[279] (2017) ont observé qu'environ 60% des LNC et 20% des PCG étaient tissus-spécifiques, ce qui est très proche des résultats obtenus par Cabili *et al*. chez l'humain. Chez la poule, Muret *et al*.[269] (2017), en utilisant du foie et du tissu adipeux, ont observé que 24% des LNC étaient exprimés dans un seul tissu, versus 4% des PCG. Nous avons pour notre part étudié la tissu-spécificité dans plus de 20 tissus des PCG et LNC chez la poule, et présentons les résultats obtenus dans l'article 1.

d) classification(s)

Comme l'expriment joliment les auteurs d'une revue de la littérature de 2015 sur les classifications possibles des LNC, celles-ci ont « *mushroomed* »† ces dernières années[280], au rythme des découvertes sur ces gènes. Cependant, les auteurs montrent que les termes utilisés dans ces classifications se chevauchent parfois les uns avec les autres, et sont souvent confus. Ainsi, citons grâce à la table 1 de cette revue les classifications basées sur la taille du LNC (qui distingue par exemple des *très longs* ARN non-codants), sur l'association avec des gènes codants (sur lesquelles nous reviendrons dans la suite – notamment dans l'article 1), l'association avec d'autres éléments de l'ADN (par exemple promoteurs ou *enhancers*), l'association avec une voie biochimique (par exemple précurseurs de miRNA – voir à ce propos l'article 1), sur la conservation d'une région ou d'une séquence, sur leur expression dans

† Manière de dire qu'elles sont apparues rapidement et en grand nombre, comme des champignons après une averse

certains états physiologiques, sur l'association à des structures subcellulaires (chromatine, corps nucléaires) ou encore sur la fonction[280]. Nous avons pour notre part adopté la classification proposée par le projet GENCODE[226], basée sur le PCG le plus proche.

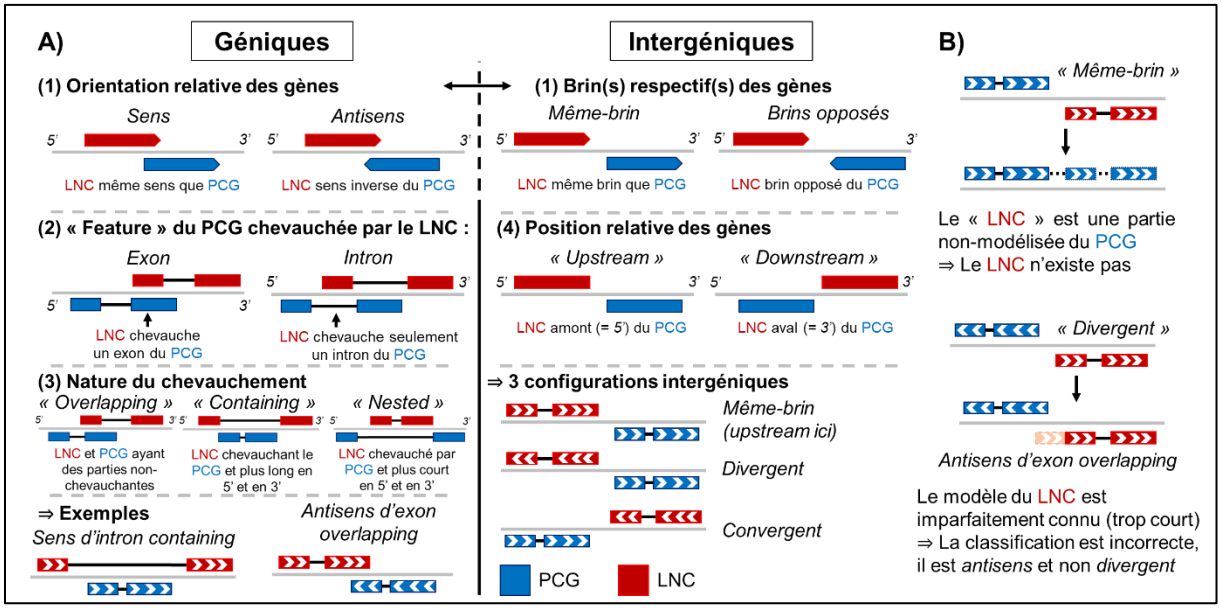


Figure 4 | Classification des LNC par rapport aux PCG. A) Illustration des situations géniques et intergéniques. Pour les géniques (*gauche*) de haut en bas, illustration de : sens et *antisens*, exonique et intronique, « *overlapping* », « *containing* » et « *nested* ». En bas, exemple d'une configuration antisens exonique et sens intronique. Pour les intergéniques (*droite*), de haut en bas : configurations même-brin et brin opposé (*haut*), *upstream* et *downstream* (*milieu*) et divergente et convergente (*bas*). **B)** Illustration de cas menant à une mauvaise classification. En haut, un LNC en même-brin d'un PCG peut en fait être une partie du PCG en question. En bas, un LNC en configuration divergente est en fait en configuration *antisens* car le LNC est mal modélisé (ici premier cas exon trop court).

Cette classification, réalisée par le module de classification de l'outil FEELnc[229], sépare les LNC comme suit (Figure 4A) :

- géniques (qui chevauchent un PCG), parmi lesquels :
  - *sens* : les deux gènes sont portés par le même brin d'ADN,
  - *antisens* : chaque gène est porté par un brin différent.

  Un LNC sens ou antisens peut enfin être exonique si le LNC chevauche tout ou partie d'un exon du PCG, intronique sinon. Enfin, selon que le LNC chevauche, est inclus ou contient le PCG, on peut préciser la configuration en *overlapping*, *containing* ou *nested*, respectivement.
- intergéniques (qui ne chevauchent pas de PCG), parmi lesquels :
  - *même-brin* : les deux gènes sont sur le même brin d'ADN, en « file indienne »,

- *divergent* : chaque gène est porté par un brin d'ADN, leur TSS sont l'un à côté de l'autre et lorsqu'ils sont transcrits, les polymérases s'éloignent,
- *convergent* : chaque gène est porté par un brin d'ADN, leur TSS sont de part et d'autre des deux gènes et lorsqu'ils sont transcrits, les polymérases se rapprochent.

Pour être plus précis, cette classification se fait au niveau des transcrits : chaque transcrit d'un LNC est classifié par rapport à chaque transcrit du ou des PCG proches (par exemple, tous les PCG dans une fenêtre de 100 kb). Ainsi, un LNC donné peut voir ses transcrits avoir différentes classifications (en termes de classe et en termes de PCG). Pour appliquer la classification à l'échelle du gène, il convient alors de décider d'un ordre de « préséance » pour les différentes classifications. Aucun travail n'a, à notre connaissance, proposé un tel ordre. Nous en avons donc proposé un dans l'article 1, en tentant de concilier (*i*) le potentiel de détecter des situations pour lesquelles les configurations, et par suite d'autres analyses, les co-expressions, amènent à faire des hypothèses biologiques et (*ii*) la capacité à détecter des situations dans lesquelles deux gènes pourraient n'en faire qu'un à cause d'une annotation incomplète, particulièrement au niveau des transcrits dont nous avons vu plus haut à quel point l'annotation était lacunaire chez la poule. Cette classification, si elle peut paraitre descriptive, n'est pourtant pas dénuée d'intérêt. En effet, différents travaux ont montré que les LNC en configuration *antisens* ou *divergente* avaient tendance à réguler l'expression du gène codant le plus proche[281,282] ou en tous cas à y être corrélé, que ce soit positivement[283–285] ou négativement[286,287], parfois à travers un promoteur bidirectionnel dans le cas des divergents[288–290].

Enfin, rappelons que notre connaissance des génomes reste globalement imparfaite. De manière générale, on peut dire que les PCG sont mieux connus que les LNC, et que les génomes humains ou d'espèces modèles comme la souris sont mieux connus que les génomes des espèces d'élevage. Ces différences sont particulièrement prégnantes au niveau des transcrits : Ensembl v101 répertorie 5 transcrits par gène chez l'humain contre 1.6 chez la poule. En détail, on compte chez l'humain 3 transcrits par LNC et 8 par PCG, et chez la poule, 1.6 transcrits par gène pour ces deux classes. Les modèles de gènes sont donc amenés à évoluer et donc les classifications aussi : un LNC en position divergente ou convergente proche d'un PCG pourra se révéler à l'avenir être en fait un *antisens* à la faveur d'une meilleure modélisation d'un des transcrits, (particulièrement au niveau du premier ou dernier exon), voire même d'un transcrit jusque-là inconnu, qui ferait que les modèles géniques seraient en fait chevauchants (voir Figure 4B). De la même manière, un LNC de type même-brin peut très bien se révéler être en fait un bout du PCG, faisant partie d'un transcrit encore non-modélisé. Ce dernier cas est

particulièrement crédible lorsque les deux gènes présentent une forte corrélation d'expression. Un tel cas a été mis en évidence par le laboratoire (Muret *et al.* en 2019[259]) : une autre équipe avait utilisé des siRNA dirigés contre les transcrits d'un LNC situé à proximité sur le génome et en même-brin que le gène *SCD*, provoquant une diminution de l'expression de *SCD*. Les deux gènes ne faisaient en fait qu'un : le LNC était un exon non-modélisé de *SCD*, les siRNA ciblaient donc en réalité les transcrits *SCD* directement.

e) Atlas des LNC

À notre connaissance, il n'existe pas de travaux qui ont pour ambition de rassembler les répertoires existants de LNC des espèces d'élevage, contrairement à l'être humain[226,291–293]. Il existe pourtant quelques bases de données, en plus des références Ensembl ou NCBI, qui recensent des LNC dans les espèces d'élevage comme ALDB[294] et NONCODE[295]. Il existe également quelques travaux visant à modéliser de nouveaux LNC même s'ils portent en général sur un faible nombre d'individus ou de tissus, ce qui ne permet pas d'établir de larges catalogues (voir les études chez le bovin[296], le mouton[297], le poulet[269], chez le poulet, le porc et le bovin[278], ou encore chez les quatre espèces, bovin, porc, chèvre et poule[102]). Dans ce contexte, nous avons étendu l'annotation de référence Ensembl du génome de la poule en agrégeant les modèles de LNC des bases de données publiques et des modèles nouvellement générés au laboratoire. Avec l'outil FEELnc, nous avons ensuite classé les LNC par rapport au PCG le plus proche et avons étudié leur expression, et notamment leur tissu-spécificité dans plus de 20 tissus (*cf*. article 1).

f) rôles et mécanismes d'action des LNC

Les LNC sont, on l'a dit, d'importants régulateurs de l'expression des gènes. Comme nous allons le voir, ils agissent aux différentes étapes de l'expression des gènes : aux niveaux pré-transcriptionnel et transcriptionnel en agissant sur la structure et l'accessibilité de la chromatine et sur la transcription elle-même, au niveau post-transcriptionnel en agissant sur la stabilité des ARN.

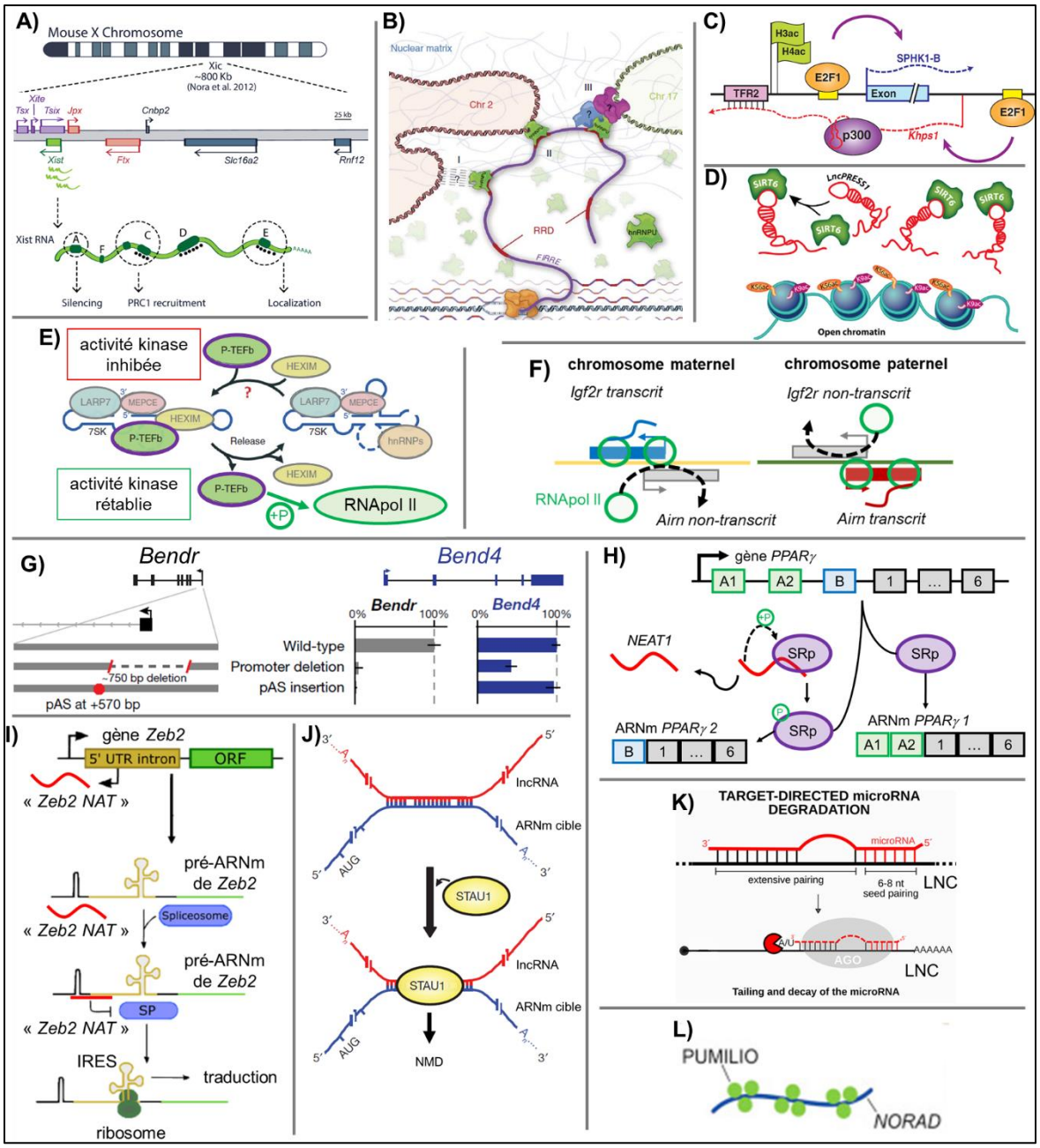


Figure 5 | Vue générale des différents mécanismes d'actions des LNC développés dans cette partie. A) *Xist* provoque la transformation d'un des deux chromosomes X des femelles mammifères en corpuscule de Barr via différents processus liés à des domaines fonctionnels du LNC. *Adapté de* [298]. **B)** *Firre* entraîne le rapprochement physique de 5 *loci*. Ici 3 *loci* du génome de la souris sont représentés : 2 sur le chromosome 2 et 1 sur le chromosome 17. *Adapté de* [299]. **C)** *Khps1*, *antisens* selon notre classification de *SPHK1*, recrute p300/CBP, provoquant une modification de la structure chromatinienne au niveau du promoteur de *SPHK1* et sa transcription. *Reproduit de* [284]. **D)** *lncPRESS1* interagit avec SIRT6, l'empêchant de retirer des marques d'histones favorisant la transcription au niveau de promoteurs de gènes impliqués dans la pluripotence dans des cellules souches embryonnaires. *Reproduit de* [300]. **E)** L'ARN *7SK* contrôle l'activité kinase du facteur l'élongation P-TEFb : lié au complexe ribonucléoprotéique dont fait partie 7SK, l'activité kinase est inhibée. Lorsqu'il est libre, P-TEFb phosphoryle des ARN polymérases II, activant la transcription. *Adapté de* [301]. **F)** L'expression du LNC *Airn*, *antisens* d'*Igfr2*, par le chromosome paternel empêche le recrutement d'ARN polymérases II et la transcription d'*Igfr2*. Le phénomène inverse s'observe sur le chromosome maternel. *Adapté de* [302]. **G)** *partie haute* : localisations respectives de *Bendr* et *Bend4*. *Partie basse* : par rapport au contrôle (« *wild-type* »), la délétion d'une région de 750 pb autour du premier exon de *Bendr*, supposée correspondre au promoteur, abolie l'expression de *Bendr* et diminue de moitié environ celle de *Bend4*. *Suite de la légende en haut de la page suivante.*

L'insertion d'un signal de poly-adénylation précoce (*pAS*) dans la séquence de *Bendr* entraine la production d'un transcrit très tronqué, difficilement détectable, mais n'a pas d'effet sur l'expression de *Bend4*. C'est donc le fait que *Bendr* soit transcrit, et non sa séquence, qui influence l'expression de *Bend4*. *Adapté de* [303]. **H)** Le LNC *NEAT1* provoque la phosphorylation d'une *SR protein* (SRp), entrainant l'épissage alternatif du gène *PPARγ*. *Adapté de* [304]. **I)** le LNC « *Zeb2 AS* », antisens de *Zeb2* selon la classification que nous avons utilisée, inhibe l'action du complexe d'épissage (*Spliceosome*, SP), entrainant la production d'un transcrit possédant un site interne d'entrée du ribosome (IRES, *Internal Ribosome Entry Site*). Ce site permet la traduction du transcrit. *Adapté de* [304]. **J)** Des LNC peuvent se lier au 3'-UTR de certains ARNm, formant de l'ARN double-brin ciblé et dégradé par STAU1 lors du *STAU1-mediated messenger RNA decay* (SMD). *Adapté de* [305]. **K)** L'appariement de microARN, comme miR-7 à certains LNC, comme *Cyrano*, provoque la dégradation du microARN via la *target RNA–directed miRNA degradation* (TDMD). *Adapté de Sophie Mockly, communication personnelle*. **L)** Le LNC *NORAD* séquestre des protéines PUMILIO1 et 2, les empêchant de favoriser la dégradation d'ARNm. *Adapté de* [306].

Structure et accessibilité de la chromatine – L'exemple probablement le plus frappant d'une modulation de la chromatine par un LNC est la transformation d'un des deux chromosomes X des femelles mammifères en corpuscule de Barr, phénomène que nous avons évoqué plus tôt à propos de l'expression mono-allélique. Chez la souris, cette transformation est due à plusieurs LNC portés par le chromosome X, dont *Xist* (*X-inactive specific transcript*, le plus connu), mais également *RepA*, *Tsix*, *Xite* ou encore *Jpx*, tous localisés dans une région appelée le *X-inactivation centre* (*Xic*)[307]. Lors de ce processus, *Xist* est exprimé par le chromosome qui va devenir le corpuscule de Barr et le « recouvre », provoquant *in fine* la répression de toute expression[308], après différentes modifications épigénétiques (méthylation de l'ADN et des histones par exemple)[307] et potentiellement d'autres mécanismes encore mal compris[309] (Figure 5A). Autre exemple, *Firre* (« *functional intergenic repeating RNA element* », Figure 5B) est supposé permettre le rapprochement physique de son site de transcription et de cinq autres *loci* (portés par les chromosomes 2 [2 *loci*], 9, 15 et 17 chez la souris). Les contacts entre *Firre* et ces 5 sites ont été mis en évidence par une méthode permettant de *cross-linker* ARN et ADN puis de séquencer l'ADN liée à un ARN d'intérêt (méthode RAP, *RNA antisens purification*)[310], et la colocalisation de *Firre* et des 5 sites mis en évidence par hybridation *in situ* en fluorescence (FISH, *fluorescence in situ hybridization*). Plus précisément, le facteur nucléaire hnRNPU se lie d'une part à ces *loci* chromosomiques et d'autre part à *Firre* via une séquence répétée[299]. Notons que ni *XIST* ni *FIRRE* n'ont d'orthologue connu chez la poule.

Concernant l'accessibilité de la chromatine, citons chez l'humain *Khps1*, un LNC antisens de *SPHK1* qui recrute l'acétyltransferase *p300/CBP*, provoquant une modification de la structure chromatinienne au niveau du promoteur de *SPHK1*, favorisant sa transcription[284] (Figure 5C). Enfin, *lncPRESS1*, interagit avec la protéine codée par *SIRT6* (une histone déacétylase) chez l'humain par un mécanisme encore non-élucidé. Cette interaction empêche SIRT6 de retirer des

marques H3K56ac et H3K9ac au niveau de promoteurs de gènes impliqués dans la pluripotence dans des cellules souches embryonnaires (Figure 5D). Cela maintient l'expression des gènes en question, et donc la capacité de ces cellules à la pluripotence[300].

Transcription – Des LNC peuvent réguler l'ARN polymérase II (et donc la transcription d'autres gènes). Ils ne tombent en revanche pas dans notre définition « stricte », puisqu'ils ne sont pas poly-adénylés et ne sont d'ailleurs pas eux-mêmes transcrits par l'ARN polymérase II mais par l'ARN polymérase III. On peut tout de même citer chez l'humain l'ARN *Alu*, un ARN non-codant de 281 nt[311] issu de l'élément transposable du même nom[312]. Cet ARN réprime la transcription après un choc thermique en se liant à l'ARN polymérase II et en s'insérant dans les complexes protéiques au niveau des promoteurs[311]. De même, chez la souris on peut citer l'ARN B2 (178 nt[313]) qui inhibe également l'action de l'ARN polymérase II après un choc thermique[314]. À l'inverse, l'ARN 7SK (330 nt[315]) contrôle l'activité kinase du facteur d'élongation P-TEFb : lorsque ce facteur est lié au complexe ribonucléoprotéique dont fait partie 7SK, son activité kinase est inhibée. Au contraire, lorsqu'il en est libéré, P-TEFb va phosphoryler des ARN polymérases II, ce qui entraine leur activation[301] (Figure 5E).
Enfin, il est également possible que la transcription d'un gène donné ne soit pas régulée par le transcrit d'un autre gène, mais plutôt par le fait que cet autre gène subisse la transcription[316]. Ainsi, chez la souris, la répression transcriptionnelle du PCG *Igfr2*, un gène sous empreinte parentale (exprimé par le chromosome maternel[317]) n'est pas liée à l'action du transcrit issu du gène LNC *Airn* (*antisens* de *Igfr2*, exprimé par le chromosome paternel), mais uniquement au fait qu'il soit transcrit : *Airn* chevauche le promoteur de *Igfr2*, et sa transcription empêche donc le recrutement de l'ARN polymérase[318] (Figure 5F). D'autres observations du même genre ont été faites chez la souris par Engreitz *et al.*[303] en 2016. Les auteurs ont mis en évidence cinq LNC et quatre PCG dont la suppression du promoteur affectait l'expression d'un gène proche. Par exemple, ils montrent que la délétion d'une région de 750 pb, incluant le premier exon du LNC *Bendr* et la région environnante, qui est supposée correspondre au promoteur, abolit l'expression de *Bendr* et diminue l'expression du PCG *Bend4*, situé en position divergente à 35kb. En revanche, l'insertion d'un signal de poly-adénylation très précoce dans le LNC *Bendr*, si elle entraine la production d'un transcrit très tronqué (et difficile à détecter), n'a aucun effet sur l'expression de *Bend4*. Ainsi, l'expression de *Bend4* dépend de la transcription de *Bendr* (c'est-à-dire du fait que le locus subisse l'action de la polymérase II), mais pas du transcrit produit (Figure 5G).

Maturation des ARN transcrits – Les *nuclear speckles* (littéralement, « mouchetures nucléaires ») sont des domaines nucléaires dépourvus de membrane et enrichis en facteurs d'épissages des pré-ARNm[319]. Avec le nucléole, les *paraspeckles* et autres corps de Cajal font partie de la famille des corps nucléaires (« *nuclear bodies* »)[320] dont tous les rôles ne sont pas encore élucidés. On trouve dans les *nuclear speckles* la majeure partie des copies de *MALAT1* (aussi appelé *NEAT2*), l'un des LNC les plus abondants du noyau cellulaire[219]. Ce LNC régule le niveau de phosphorylation de facteurs d'épissage (*Serine/Arginine splicing factors*, ou *SR proteins*), ce qui en retour module leur activité et le niveau d'épissage alternatif observé[321]. Notons cependant que les rôles de *MALAT1* dans le noyau (y compris hors des *nuclear speckles*) semblent plus variés et sont encore relativement flous[219]. De même, *NEAT1* un autre LNC très abondant, présent cette fois dans les *paraspeckles*, s'associe avec une *SR protein* et influence l'épissage de *PPARγ* en deux isoformes durant la différenciation des adipocytes[304] (Figure 5H). Il est intéressant de constater que malgré leurs rôles apparemment clefs, en tous cas chez les mammifères, ni *NEAT1* ni *MALAT1* (alias *NEAT2*) ne sont modélisés chez la poule. En revanche, la séquence en 3' de *MALAT1*, qui forme une triple hélice d'ARN, se retrouve dans de nombreux *loci* chez les vertébrés[322], formant une classe de LNC « *MALAT1-like* ». Ceci nous amène à nouveau à nous interroger sur l'importance, chez les LNC, de la structure primaire *versus* les structures secondaires, tertiaires, ou de petits patches de structure primaire.
Enfin, le LNC « *Zeb2 NAT* » situé en *antisens* du gène *Zeb2* empêche la reconnaissance d'un site d'épissage de *Zeb2* en s'y liant via une interaction ARN:ARN. L'épissage n'étant pas réalisé, le transcrit non-épissé de *Zeb2* conserve un « *internal ribosome entry site* » (*IRES*) qui lui permet d'être traduit (Figure 5I), ce qui induit la transition épithélio-mésenchymateuse[285], un phénomène intervenant aussi bien dans des processus biologiques normaux (développement embryonnaire, réparation tissulaire) que pathologiques (métastases)[323].

Stabilité des ARN transcrits – Quelques LNC ont été montrés comme influençant la stabilité d'ARNm, comme par exemple *EGFR-AS1* (en antisens de *EGFR*) qui semble stabiliser *EGFR* en s'y liant sur ~200 nt[324]. A l'inverse, certains LNC promeuvent la dégradation de leur cible, par exemple en se liant au 3'-UTR de certains ARNm, formant localement de l'ARN double-brin qui est ciblé et dégradé par *STAU1* dans le cadre du *STAU1-mediated messenger RNA decay* (*SMD*)[305] (Figure 5J).
En revanche, les LNC semblent plus étudiés pour leurs interactions avec les miRNA. En effet, une hypothèse[325] propose que les LNC puissent agir comme des « éponges à miRNA » dans la cellule, c'est-à-dire qu'ils seraient à même de « titrer » les miRNA présents, faisant en quelque

sorte concurrence aux ARNm. Deux mécanismes sont proposés quant au devenir des LNC et miRNA : dans le premier, les LNC captant les miRNA seraient dégradés suivant le processus classique impliquant le complexe RISC[326]. Cette possibilité se heurte cependant à l'objection que la quantité de LNC cibles présents dans la cellule n'est pas suffisante pour obtenir un effet biologiquement pertinent[327]. Un mécanisme légèrement différent a été proposé par Kleaveland *et al.*[328] en 2018. Dans leur article, les auteurs montrent notamment chez la souris que miR-7 s'apparie au LNC *Cyrano*, mais que cet appariement, au lieu de provoquer la dégradation des deux transcrits (le LNC et le miRNA), ne provoque que celle du miRNA via la *target RNA–directed miRNA degradation* (TDMD), qui consiste en l'ajout de nucléotides en 3' du miRNA suivi de sa dégradation[329,330] (Figure 5K). Pour finir, un LNC qui semble en effet agir en séquestrant des éléments dégradant les ARNm est le LNC *NORAD*[331]. Ce LNC, majoritairement présent dans le cytoplasme, contient environ une vingtaine de sites reconnus par les enzymes PUMILIO 1 et 2, enzymes qui reconnaissent ces sites sur les ARNm et stimulent la déadénylation et le décoiffage des transcrits[332] (Figure 5L). En séquestrant ces protéines, *NORAD* ajoute une couche de complexité dans la régulation de la quantité d'ARNm dans les cellules.

IV – Objectifs de la thèse

1. Étude de la composante génétique de l'efficience alimentaire

La thèse avait trois objectifs concernant l'efficience alimentaire : (*i*) étudier les gènes impliqués dans l'efficience alimentaire, (*ii*) rechercher des gènes causaux de l'efficience alimentaire et (*iii*) étudier les possibles interactions entre l'efficience alimentaire (EA) et le régime, en soumettant deux lignées de poules divergentes pour l'EA à un régime hypo-énergétique. Dans les 3 objectifs, nous nous sommes concentrés à chaque fois sur l'expression des gènes, codants ou longs non-codants.

a) modèle d'étude : les lignées R+ et R- et leurs croisements réciproques

Notre modèle d'étude consiste principalement en deux lignées expérimentales de poules pondeuses, divergentes pour l'efficience alimentaire (précisément, pour la *Residual Feed Intake*, RFI, présentée en introduction), appelées R+ et R-[333]. Les premières sont « inefficientes alimentairement » : elles ont une RFI positive, donc leur consommation alimentaire observée est plus élevée que celle prédite par le modèle présenté en début d'introduction. Les R- à l'inverse sont « efficientes alimentairement » : elles ont une RFI négative, donc leur consommation alimentaire observée est plus faible que celle prédite par le modèle. Ces divergences de RFI sont la résultante d'une sélection divergente entreprise depuis 1976, soit environ 40 ans (donc environ 40 générations) sur la RFI, sur des animaux des deux sexes (avec un modèle différent pour les mâles, ne prenant pas en compte la ponte). Cette divergence d'efficience alimentaire résiduelle entre les deux lignées R+ et R- est accompagnée par d'autres variations, parfois très fortes, et pour certaines associées génétiquement à la RFI (par exemple, la longueur des barbillons, corrélation génétique r_g = 0.19[334]). Ainsi les lignées R+ et R- sont fortement divergentes pour le FCR (*Feed Conversion Ratio*, 3.6 pour les R+ contre 2.3 pour les R-, r_g = 0.38[334]) et pour la prise alimentaire (r_g = 0.40[334]), caractères fortement corrélés génétiquement à la RFI. Des différences de température corporelle ont également été observées à l'état nourri chez les mâles (plus élevée chez les R+[335]), mais ce caractère n'est pas corrélé génétiquement au RFI (r_g = 0.04[334]). Les lignées sont également fortement contrastées pour la masse de gras abdominal, plus importante chez les R- que les R+ (environ 7% du poids du corps chez les R- contre 5% chez les R+[336]), et dont la corrélation génétique avec la RFI est à ce jour inconnue. En ce qui concerne le gras abdominal, Tixier *et al*.[337] (1988) ont par ailleurs observé entre des mâles des deux lignées, âgés de 8 semaines, que la divergence de la masse de gras

abdominal ramenée au poids du corps était observable avant la divergence de FCR, calculé comme le ratio entre la masse d'aliment consommé et le gain de poids entre 5 et 8 semaines. Dans la suite, nous aurons tendance par raccourci à dire que nos deux lignées sont « divergentes pour la RFI », voire même « divergentes pour l'efficience alimentaire ». Il faudra bien sûr comprendre par-là « et également divergentes pour les caractères associés, tels que l'adiposité corporelle ». Nous avons étudié dans ces lignées les transcriptomes par RNA-seq de femelles R+ et R- âgées de 31 semaines dans quatre tissus : le tissu adipeux, lieu du stockage et de la mobilisation des acides-gras qui composent les réserves énergétiques des animaux ; le foie, lieu notamment de la synthèse des acides-gras ; l'hypothalamus, centre de régulation de l'homéostasie énergétique ; et le sang, tissu circulant qui assure le transport des nutriments et des messagers moléculaires dans tout l'organisme. Ces études transcriptomiques ont été faites en relation avec d'autres types de données selon l'objectif travaillé : données phénotypiques (production, qualité), lipidomiques et métabolomiques, traces de sélections (régions du génome dans lesquelles la différence des fréquences alléliques observées entre les deux lignées n'est pas dû au hasard mais à l'action de la sélection divergente). En complément des F_0, des animaux hybrides F_1 de parents R+ et R- ont été générés afin de mettre en évidence des gènes *cis*-régulés en étudiant l'expression allèle-spécifique chez ces animaux. Ces animaux F_1 sont issus de croisements réciproques, c'est-à-dire que la moitié des animaux F_1 ont des pères R+ et des mères R-, l'autre moitié des pères R- et des mères R+. Ainsi, aux recombinaisons près, ces animaux F_1 portent un chromosome d'origine R+ et un chromosome d'origine R-. Les variants régulateurs présents dans les génomes des deux lignées devraient donc être à l'état hétérozygote chez un certain nombre d'individus F_1 dans la mesure où chaque trace de sélection a, par définition, un haplotype fixé ou quasi-fixé dans une des deux lignées.

b) étude des gènes impliqués dans l'efficience alimentaire et recherche de gènes causaux de ce caractère

Nous avons cherché à mieux comprendre les mécanismes à l'œuvre dans le foie et le tissu adipeux et pouvant expliquer la variation de l'efficience alimentaire et des caractères associés entre nos deux lignées. Pour ce faire, nous avons comparé dans ces tissus l'expression des gènes, aussi bien PCG que LNC grâce à l'annotation du génome de la poule enrichie en LNC que nous avons générée (voir les objectifs liés à l'annotation fonctionnelles des génomes ci-après), et étudié les processus biologiques associés aux PCG différentiellement exprimés. Nous avons ensuite comparé la composition du lipidome dans ces deux tissus, et étudié conjointement dans le foie le transcriptome, le lipidome et le métabolome, à l'aide d'une méthode intégrative

basée sur l'analyse factorielle multiple (AFM) et l'analyse des co-expressions. Grâce à notre annotation du génome enrichie, nous avons là encore pu nous intéresser aux rôles éventuels des LNC.

Ensuite, nous avons cherché des gènes potentiellement causaux de la variation de ce caractère. Pour ce faire, nous avons combiné trois approches. D'abord, des traces de sélection, qui sont des régions du génome dans lesquelles la différence des fréquences alléliques entre les lignées est dû à la sélection et non uniquement à la dérive génétique, qui ont été détectées par une autre équipe. Dans ces régions sont censés se trouver les variants causaux de la différence d'efficience alimentaire. Nous y avons donc d'abord recherché les variants dans les régions codantes affectant la fonction d'une protéine. Ensuite, nous avons identifié dans et à proximité des traces de sélections les gènes, PCG ou LNC, différentiellement exprimés entre les deux lignées (grâce au travail précédent). Parmi ces gènes différentiellement exprimés, nous avons recherché ceux qui le sont à cause d'une *cis*-régulation, ce qui en fait donc de bons candidats causaux de la différence d'EA. Pour détecter ces derniers, nous avons étudié l'expression allèle-spécifique dans la F_1 issue du croisement réciproque de parents de chaque lignée.

c) étude des gènes impliqués dans l'adaptation à un aliment hypo-énergétique

Pour détecter une éventuelle interaction entre lignée et régime hypo-énergétique, et étudier les mécanismes impliqués dans l'adaptation R+ et R- à un aliment hypo-énergétique, nous avons comparé les transcriptomes des quatre tissus cités plus haut chez des animaux des deux lignées, nourris pendant 14 semaines avec un aliment hypo-énergétique ou un régime contrôle, en lien avec la comparaison de leurs performances.

2. Annotation fonctionnelle du génome de la poule par RNA-seq

Dans cette seconde partie, nous nous sommes fixé trois objectifs, utilisant les données RNA-seq générées dans la partie précédente. En retour, les résultats de cette seconde partie ont été utilisés pour mieux comprendre l'efficience alimentaire. D'abord, nous avons cherché à étendre notre connaissance du génome de la poule en termes de gènes d'ARN longs non-codants (LNC), et à fournir pour ces LNC des informations liées à leurs positions et à leurs expressions tissulaires. Nous avons souvent utilisé ces informations dans le cadre de l'étude des gènes et tissus impliqués dans l'efficience alimentaire et la recherche de gènes causaux de ce caractère. Ensuite, nous avons exploré le potentiel du RNA-seq pour la détection de variants génomiques de type SNP. Enfin, et en lien avec le point précédent, nous avons mis au point un pipeline permettant l'étude de l'expression allèle spécifique dans plusieurs tissus et à travers plusieurs individus en nous basant sur phASER et ses déclinaisons, ce qui était un préliminaire nécessaire à la recherche de gènes causaux de l'efficience alimentaire.

a) extension du catalogue de référence Ensembl en LNC chez la poule

Nous avons étendu l'annotation du génome de la poule en LNC en intégrant à l'annotation de référence Ensembl (v94, octobre 2018) des modèles de LNC (*i*) générés au laboratoire à partir de 364 échantillons de RNA-seq issus de trois tissus (tissu adipeux, sang et foie), et (*ii*) issus de bases de données publiques (NONCODE, NCBI, ALDB et FR-AgENCODE). Ensuite, nous avons étudié l'ensemble de ces LNC sous différents aspects, et notamment leur expression dans 21 tissus, grâce à des données publiques et dans 5 tissus grâce à des données disponibles au laboratoire. L'annotation ainsi générée a été utilisée dans le cadre de l'étude des mécanismes tissulaires de l'efficience alimentaire et de la recherche de gènes causaux de ce caractère.

b) détection des variants de type SNP par RNA-seq

Pour évaluer le potentiel de détection de SNP à l'aide de données de RNA-seq, ce qui est notamment nécessaire pour l'étude de l'expression allèle-spécifique (*cf.* paragraphe suivant et partie « étude des mécanismes tissulaires de l'efficience alimentaire et recherche de gènes causaux de ce caractère » dans les objectifs liés à l'efficience alimentaire), nous avons comparé les SNP détectés par RNA-seq à ceux détectés par DNA-seq dans deux populations de poules pour lesquelles les deux types de données étaient disponibles sur les mêmes échantillons.

Nous sommes allés plus loin en étudiant le potentiel de détection par RNA-seq, non plus de SNP dans une population, mais de génotypes dans des individus, et avons proposé des critères

permettant d'assurer une bonne concordance entre les génotypes détectés en RNA-seq et ceux détectés en DNA-seq. Enfin, nous avons présenté quelques applications possibles de la détection de SNP par RNA-seq, et notamment l'étude de l'expression allèle-spécifique, que nous avons mis à profit pour la recherche de gènes causaux de l'efficience alimentaire et développée dans le point suivant.

c) mise au point d'un pipeline d'analyse de l'expression allèle-spécifique reposant sur phASER et ses déclinaisons

Pour permettre la recherche de gènes causaux de l'efficience alimentaire, l'analyse de l'expression allèle-spécifique (ASE) dans les traces de sélection à travers plusieurs individus et tissus est une étape incontournable. Elle implique la détection de variants par RNA-seq (*cf.* paragraphe précédent) puis l'analyse de l'ASE dans plusieurs tissus et individus. Dans la mesure où phASER étudie l'ASE par échantillon (tissu × individu) et n'agrège les résultats à l'échelle de la population étudiée que dans des cas très particuliers, nous avons mis au point un pipeline reposant sur phASER et des scripts *ad hoc* permettant l'étude de l'ASE dans notre F_1.

Nous avons vu que les résultats de la partie visant à améliorer l'annotation fonctionnelle du génome de la poule grâce au RNA-seq, dont nous venons de présenter les objectifs, ont été largement utilisés pour l'étude de la composante génétique de l'efficience alimentaire. En particulier, la partie « recherche de gènes causaux de l'efficience alimentaire » s'appuie sur cette annotation fonctionnelle que ce soit pour des aspects techniques (mise en place du pipeline d'ASE qui repose sur la détection de SNP par RNA-seq) ou biologiques (recherche de LNC potentiellement causaux). Elle s'appuie également sur les listes de gènes impliqués dans l'efficience alimentaire, à savoir les gènes différentiellement exprimés entre les deux lignées.

Ainsi, nous présenterons les résultats obtenus durant la présente thèse dans l'ordre suivant :

I. d'abord, les résultats des objectifs liés à l'annotation fonctionnelle du génome de la poule grâce au RNA-seq,
II. ensuite, deux résultats des objectifs liés à l'étude de la composante génétique de l'efficience alimentaire : étude des gènes impliqués dans l'efficience alimentaire et des gènes impliqués dans l'adaptation à un aliment hypo-énergétique,
III. enfin, la recherche de gènes causaux de l'efficience alimentaire.

Articles et travaux complémentaires

I – Annotation du génome de la poule en ARN longs non-codants et détection de SNP par RNA-seq pour la mise en place d'une démarche d'analyse de l'expression allèle-spécifique

Avant de nous plonger dans la compréhension des gènes et tissus sous-jacents à la différence d'efficience alimentaire entre nos lignées modèles (partie II) et dans la recherche de gènes candidats causaux (partie III) nous avons cherché à étendre notre connaissance du génome de la poule en termes de gènes d'ARN longs non-codants (LNC), d'importants régulateurs de l'expression (§ 1). Ensuite, nous avons exploré le potentiel du RNA-seq pour la détection de variants génomique de type SNP (§ 2), étape clef pour l'analyse de l'expression allèle-spécifique (ASE), analyse qui a nécessité la mise en place d'un pipeline idoine (§ 3).

1. Un atlas intégratif des gènes à ARN longs non-codants et leur annotation à travers 25 tissus (article 1)

a) contexte et objectifs

Les ARN longs non-codants (LNC) sont impliqués dans de nombreux processus biologiques, et régulent l'expression d'autres gènes (codants des protéines ou non) à toutes les étapes de l'expression génique. Cependant, étant très peu exprimés – environ 10 fois moins que les gènes codant de protéines (PCG) – ces gènes sont moins connus que les PCG. En effet, s'ils sont assez bien répertoriés chez l'humain et les espèces modèles – 15 000 LNC pour 20 000 PCG chez l'humain, 10 000 LNC pour 20 000 PCG chez la souris ; chez les autres espèces en revanche, comme les espèces d'élevage, leur catalogue est bien plus lacunaire : alors qu'environ 20 000 PCG sont également répertoriés chez la poule, la vache et le porc, on compte environ 4 500 chez la première, aucun chez le second, et moins de 500 chez le dernier (Ensembl, version 94 – octobre 2018, assemblage du génome *Gallus_gallus-5.0*). La dernière version de l'annotation Ensembl (v101 – août 2020) qui correspond au nouvel assemblage du génome de la poule sorti en janvier 2019 (GRCg6a) compte 5 506 LNC, nombre toujours bien en deçà de l'annotation humaine. Cela étant, connaître l'existence d'un LNC ne permet pas d'en connaître la fonction, et les rôles de ces gènes sont globalement très mal connus : chez l'humain, moins de 1% des LNC ont une fonction qui leur est associée[338]. En effet, il n'existe pas d'outils permettant de

prédire une fonction par analyse de séquence, contrairement aux PCG. D'une part, il n'est pas possible de prédire une fonction en se basant sur la similarité de séquence par recherche de motifs spécifiques de fonction dans la protéine associée[339,340]. Par ailleurs, comme indiqué en introduction, les LNC sont mal conservés en séquence entre espèces, particulièrement pour des espèces éloignées évolutivement et il est donc difficile d'annoter fonctionnellement un LNC par recherche d'orthologues comme cela se fait souvent pour les PCG. Nous nous proposons ici de poursuivre un travail engagé au cours d'une thèse précédente[177] : l'extension de l'annotation du génome de la poule en modèle de LNC. Une fois cette extension réalisée, nous avons tenté d'annoter les LNC de différentes manières, notamment en utilisant le gène codant le plus proche, ou bien grâce à leur expression dans différents tissus. Ce travail a été réalisé dans le cadre d'une collaboration avec Maria Bernard et Christophe Klopp de l'équipe transversale de bio-informatique Sigenae (Toulouse).

b) matériels et démarche

Lors de ce travail, nous avons dans un premier temps étendu le catalogue des LNC connus chez la poule dans la base de données Ensembl (v94), puis avons, par différentes approches, tenté d'apporter quelques annotations fonctionnelles les caractérisant en étudiant leur expression dans plus de 20 tissus. Pour cela, nous avons utilisé des données disponibles au laboratoire obtenus dans le cadre de différents projets, soit français (*SOSrnaSEQ*, INRA 2010 ; *Elastic*, INRA 2013 ; *FatInteger*, ANR 2012–2015 ; *Chickstress*, ANR 2014–2018) soit européens (*Feed-a-Gene*, 2015–2020) et des données disponibles publiquement. L'ensemble de ces données a été traité en collaboration avec la plateforme Sigenae (Toulouse).

Pour l'extension du catalogue, nous sommes partis des LNC connus dans la base de données Ensembl (assemblage *gallus_gallus_5*, version 94), qui nous a servi de référence. À ces LNC, nous avons ajouté des gènes modélisés dans l'équipe à partir de 364 librairies RNA-seq (avec environ 40 millions de *reads* pairés de 150 pb chacune) issues de 3 tissus (tissu adipeux, sang, foie), après un filtrage sur l'expression (voir Figure 1a de l'article), avant d'ajouter des modèles de gènes issus de trois bases de données publiques, dont les modèles générés dans le cadre du projet FR-AgENCODE[102], auquel le laboratoire a participé. Nous avons adopté une stratégie consistant à agréger séquentiellement ces sources, n'ajoutant à chaque étape que les modèles de gènes dont aucun exon ne se superposait sur le même brin à un exon d'un modèle déjà présent. L'ordre d'ajout des bases de données a été déterminé à l'aide d'un proxy de la qualité

de modélisation des extrémités 5' des transcrits sur la base des données CAGE du consortium FANTOM[341,342].

Une fois l'extension du catalogue Ensembl réalisée, nous avons suivi plusieurs approches pour tenter d'annoter ces LNC. D'abord, nous avons classifié chaque LNC grâce à sa configuration par rapport au PCG le plus proche. En effet, comme nous l'avons vu en introduction, certaines configurations pour ces paires LNC:PCG peuvent suggérer une implication dans un même processus biologique, qui est renforcée si les deux membres sont en plus co-exprimés. Citons par exemple la configuration divergente : LNC et PCG chacun sur un brin, sans superposition, avec des sites d'initiation de transcription contiguës, leur transcription se faisant donc en directions opposées, et pouvant donc initier une co-expression – voir aussi Figure 2d de l'article ci-après ou la Figure 4 de l'introduction de la présente thèse. Ensuite, nous avons étudié la tissu-spécificité de l'expression des LNC, c'est-à-dire la tendance à être exprimé dans un ou quelques tissus, et très peu dans les autres, voire même dans un seul tissu. Ici, l'hypothèse est qu'un gène tissu-spécifique a probablement une fonction en lien avec le ou les tissus dans le(s)quel(s) il est exprimé. Cette étude de tissu-spécificité a été réalisée à l'aide de la métrique τ, un score de spécificité qui, pour chaque gène, compare son expression par tissu à son expression la plus élevée, et qui prend ses valeurs entre 0 (gène parfaitement ubiquiste, exprimé dans tous les tissus au même niveau) et 1 (gène exprimé dans un unique tissu).

c) résultats

La première étape d'extension du catalogue de référence Ensembl v94 nous a permis de générer un fichier d'annotation répertoriant 30 084 LNC, soit une augmentation d'un facteur 6.5 par rapport au catalogue Ensembl seul (Figure 1c de l'article), parmi lesquels 59% et 41% ont une expression soutenue (≥ 0.5 TPM) voire très soutenue (≥ 1TPM), respectivement. Nous avons ensuite classifié ces LNC par rapport au PCG le plus proche. Comme nous l'avons vu en introduction, puisque cette classification se fait au niveau des transcrits, il a fallu décider d'un ordre de préséance pour faire la classification à l'échelle des gènes, que nous détaillons dans le point ***f)*** ci-après.

Nous avons observé que 80% des LNC étaient intergéniques (c'est-à-dire qu'ils ne se superposent pas à un PCG sur aucun des 2 brins) dont 27% en position divergente, les 20% restants étant géniques. Ces proportions sont similaires aux résultats obtenus chez l'humain ou les autres espèces d'élevage.

En exploitant des données publiques d'expression sur 21 tissus, et en répétant nos résultats à l'aide de données d'expression sur 5 tissus disponibles au laboratoire, nous avons pu étudier les co-expression des membres de chaque type de paires entre ces tissus. Nous avons observé un enrichissement des LNC:PCG co-exprimés pour les paires divergentes (14%) et même-brins (32%) par rapport aux paires convergentes (7%, Figure 3a de l'article). Ces enrichissements peuvent être expliqués dans le premier cas par une régulation commune via un promoteur bi-directionnel[288,289,343] et probablement une fonction commune. Le second cas rend attentif à de potentielles erreurs de modélisation, qui masquent le fait que le LNC et le PCG sont en fait un même gène, d'où l'enrichissement de corrélation d'expression au travers des tissus. Enfin, nous avons observé que les corrélations négatives étaient rares au sein de ces paires (1.2%), ce qui avait aussi été observé chez l'humain[226] et le chien[270] dans des études portant sur 8 et 11 tissus, respectivement.

Nous avons constaté que 16% des micro-ARN (MIR) du génome de la poule étaient présents dans un modèle de LNC, pour 17.5% chez l'humain. Parmi ces MIR, la moitié environ (74 sur 144) était sur le même brin que le LNC. Ces LNC hôtes servent donc à priori de précurseurs pour le MIR, et peuvent ensuite avoir des fonctions en lien avec celles du MIR. Pour l'autre moitié, la transcription est initiée par un promoteur différent (puisque les deux gènes ne sont pas sur le même brin), et il est plus difficile de faire l'hypothèse d'une fonction commune. Nous sommes allés plus loin dans l'analyse en étudiant conjointement les fonctions des MIR et des PCG corrélés au LNC hôte du MIR. En partant des 185 LNC hôtes de MIR, nous n'avons retenu que ceux exprimés dans au moins un tissu (126 hôtes), et pour lesquels au moins 75% des PCG corrélés (corrélation de Spearman ≥ 0.8) avec l'hôte avaient un identifiant HGNC, afin de pouvoir réaliser une analyse d'enrichissement de termes fonctionnels. Nous n'avons ainsi retenu que 28 LNC hôtes. Ensuite, nous avons éliminé les LNC hôtes accueillant des MIR inconnus ou peu connus dans la littérature (après une recherche manuelle sur PubMed), pour pouvoir étudier conjointement le rôle supposé du MIR et les fonctions enrichies dans les PCG corrélés au LNC hôte. Nous avons ainsi obtenu 10 LNC accueillant un MIR étudié dans la littérature et avons trouvé deux LNC parmi ces 10 conservés entre la poule, l'homme et la souris, et qui semblent partager la même fonction avec le MIR hébergé. Un exemple autour d'un de ces deux LNC, orthologue chez la poule de *MIR155HG* (HG pour *Host Gene*) humain est développé en Figure 4 de l'article : nous avons constaté que les PCG les plus co-exprimés avec le LNC (sans considération de paires cette-fois) étaient enrichis en gènes associés avec les mêmes fonctions biologiques que le MIR, à savoir ici l'immunité. Ce LNC était exprimé dans des tissus ayant un rôle immunitaire, tout comme le MIR qu'il accueille, dont nous avons étudié

l'expression à l'aide de données publiques. Ces résultats nous suggèrent donc que le LNC et le MIR qu'il accueille ont des fonctions communes. Nous avons également analysé la tissu-spécificité de l'expression des LNC, c'est-à-dire leur tendance à être exprimés dans un ou quelques tissus, et à ne pas l'être ou bien à un très faible niveau dans les autres. Un tel profil d'expression suggère en effet que le LNC joue un rôle en lien avec les fonctions du tissu ou du groupe de tissus en question, et possiblement le maintien de leur différenciation. Ce travail nous a permis de confirmer que les LNC étaient plus tissu-spécifiques que les PCG (25.2% pour les LNC contre 9.8% pour les PCG ; Figure 5a de l'article) ce qui est en accord avec la bibliographie chez l'humain.

Enfin, en croisant la classification des paires avec les informations de tissu-spécificité, nous avons étudié les paires LNC:PCG dont les 2 membres étaient tissu-spécifiques. Nous avons mis en évidence le fait que les PCG des couples en position divergente ou *antisens* et dont les membres sont tissus-spécifiques dans le même tissu tendent à être des facteurs de transcription, ce qui pose la question de leur(s) rôle(s) dans le tissu en question. En effet, 14 des 45 couples LNC:PCG concernés sont impliqués dans le développement embryonnaire ou la différenciation des tissus dans lesquels ils sont exprimés (Figure 6 de l'article).

d) discussion et conclusion

Ce travail a ainsi permis d'étendre l'annotation en LNC du génome de la poule, de caractériser leur expression à travers plus de 20 tissus, et de suggérer des rôles pour certains LNC du catalogue. Il pourra être utile pour la communauté de chercheurs travaillant sur l'espèce poule, mais aussi aux chercheurs souhaitant étudier la conservation d'un LNC à travers l'évolution (humains et poules ont par exemple divergé il y a 300 millions d'années). Enfin, dans le cadre de la thèse, ce catalogue sera utilisé pour les travaux portant sur l'efficience alimentaire.

e) article publié[344]

Cet article a fait l'objet :

- d'une publication dans *Scientific Reports* : **Jehl, F.**, Muret, K., Bernard, M. et al. An integrative atlas of chicken long non-coding genes and their annotations across 25 tissues. Scientific Reports. 10, 20457 (2020). https://doi.org/10.1038/s41598-020-77586, reproduite ci-après ;

- d'une communication orale : **Frédéric Jehl**, Kévin Muret, Maria Bernard, Diane Esquerré, Hervé Acloque, Elisabetta Giuffra, Sarah Djebali, Sylvain Foissac, Thomas Derrien, Tatiana Zerjal, Christophe Klopp and Sandrine Lagarrigue. An atlas of chicken long non-coding RNAs gathering multiple sources: gene models and expression across more than twenty tissues. International Plant & Animal Genome XXVII (PAG),
  San Diego (États-Unis), le 13 janvier 2019 ;

- d'un poster : **Frédéric Jehl**, Kévin Muret, Maria Bernard, Diane Esquerré, Hervé Acloque, Elisabetta Giuffra, Sarah Djebali, Sylvain Foissac, Thomas Derrien, Tatiana Zerjal, Christophe Klopp and Sandrine Lagarrigue. An atlas of chicken long non-coding RNAs gathering multiple sources and expression across more than twenty tissues. International Plant & Animal Genome XXVII (PAG),
  San Diego (États-Unis), le 14 janvier 2019.

scientific reports

Check for updates

OPEN

An integrative atlas of chicken long non-coding genes and their annotations across 25 tissues

Frédéric Jehl[1,8], Kévin Muret[1,8], Maria Bernard[2,8], Morgane Boutin[1], Laetitia Lagoutte[1], Colette Désert[1], Patrice Dehais[2], Diane Esquerré[3], Hervé Acloque[4], Elisabetta Giuffra[5], Sarah Djebali[6], Sylvain Foissac[4], Thomas Derrien[7], Frédérique Pitel[4], Tatiana Zerjal[5], Christophe Klopp[2]✉ & Sandrine Lagarrigue[1]✉

Long non-coding RNAs (LNC) regulate numerous biological processes. In contrast to human, the identification of LNC in farm species, like chicken, is still lacunar. We propose a catalogue of 52,075 chicken genes enriched in LNC (http://www.fragencode.org/), built from the Ensembl reference extended using novel LNC modelled here from 364 RNA-seq and LNC from four public databases. The Ensembl reference grew from 4,643 to 30,084 LNC, of which 59% and 41% with expression ≥ 0.5 and ≥ 1 TPM respectively. Characterization of these LNC relatively to the closest protein coding genes (PCG) revealed that 79% of LNC are in intergenic regions, as in other species. Expression analysis across 25 tissues revealed an enrichment of co-expressed LNC:PCG pairs, suggesting co-regulation and/or co-function. As expected LNC were more tissue-specific than PCG (25% vs. 10%). Similarly to human, 16% of chicken LNC hosted one or more miRNA. We highlighted a new chicken LNC, hosting miR155, conserved in human, highly expressed in immune tissues like miR155, and correlated with immunity-related PCG in both species. Among LNC:PCG pairs tissue-specific in the same tissue, we revealed an enrichment of divergent pairs with the PCG coding transcription factors, as for example LHX5, HXD3 and TBX4, in both human and chicken.

Since their description at the end of the 1990s[1,2], long non-coding RNAs (LNC) have been shown to be involved in numerous biological processes, both at the cellular level (such as cell proliferation[3] and differentiation[4], metabolism[5] and apoptosis[6]) and at the organism level, whether influencing sex differentiation[7], agronomical traits[8,9], disease[10] or cancer[11]. LNC genes act through regulatory actions at every levels of gene expression, from the epigenetic modification of the chromatin[12] to the regulation of transcription[13] and protein translation[14].

While it is well known that complex traits are mostly driven by variants located outside the coding regions[15] that likely are regulatory variants modulating gene expression, the molecular chain of causality linking the DNA variant to the gene to the phenotype remains scarce. A better knowledge of LNC genes and of their regulatory functions could therefore be of help to identify their target genes, and possibly link them with the phenotype of interest. In model species such as human and mouse, LNC are well characterized with 15,137 and 10,177 LNC genes, respectively in Ensembl database (*versus* 20,465 and 22,604 protein coding-genes, respectively)[106]; although these estimates are nevertheless bound to increase[16] in the continuity of works like the GENCODE v7 annotation of human LNC that manually annotated 9277 LNC in 2012[17], while the human NONCODE (v5) integrated more than 96,000 LNC at the end of 2017[18]. In non-model species, such as farm species, LNC knowledge is largely incomplete. In the Ensembl release 94 reference database used in the present paper (October 2018), 4,643 LNC genes were known in chicken, none in cow and 361 in pig, *versus* 18,346; 19,994 and 22,452 PCG, respectively[16]. LNC annotation is one of the priorities of the Functional Annotation of ANimal Genomes (FAANG) initiative[19,20]. We and others recently reported LNC transcript sets for cattle, sheep, horse, pig, goat[20,21] and chicken[22].

[1]PEGASE UMR 1348, INRA, AGROCAMPUS OUEST, 35590 Saint-Gilles, France. [2]SIGENAE Platform, INRA, 31326 Castanet-Tolosan, France. [3]GENOTOUL Platform, INRA, 31326 Castanet-Tolosan, France. [4]GenPhySE UMR 1388, INRA, INPT, ENVT, Université de Toulouse, 31326 Castanet-Tolosan, France. [5]GABI UMR 1313, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France. [6]IRSD, Université de Toulouse, INSERM, INRA, ENVT, UPS, Toulouse, France. [7]IGDR UMR 6290, Univ Rennes, CNRS, 35000 Rennes, France. [8]These authors have contributed equally: Frédéric Jehl, Kévin Muret and Maria Bernard. ✉email: christophe.klopp@inrae.fr; sandrine.lagarrigue@agrocampus-ouest.fr

Among farm species, chicken represents an interesting species, both for fundamental and applied research. It is used in evolutionary studies to evaluate the level of conservation of genomics feature among species[23] and is a valuable model in the developmental biology research field[24]. Moreover, chicken is a species of great economic value, being one of the most consumed in the world, and representing a 300 billion dollars market[25].

The first aim of this study was to enrich the chicken Ensembl gene catalogue in long non-coding genes. For this we used new LNC genes that were computationally predicted in this study using 364 RNA-seq samples, as well as LNC genes available in the public databases NCBI[26], NONCODE[27] ALDB[28] and FR-AgENCODE[21,29]. This enriched gene catalogue can be useful for the scientific community working on chicken, especially for those aiming at analysing gene expression rather than modelling genes, and therefore use a reference annotation such as Ensembl or NCBI at the gene level.

The second aim of our study was the characterization of the identified LNC. We analysed their expression profiles in 25 distinct tissue types. Since LNC genes are generally very weakly expressed compared to the PCG[17,30], we provide the level of expression for each LNC in the tissue in which it is the most expressed, giving an indication of how easy it is to study experimentally. Since the configuration of a LNC and its close PCG (hereafter, LNC:PCG pairs), such as the close divergent and antisense, can be an indicator of a regulatory role of the LNC on the PCG[31–34], and therefore of an involvement in a common biological function[35], we classified the LNC based on their genomic configuration with respect to their closest protein-coding gene (PCG) and screened for LNC:PCG pairs showing a significant co-expression. This latter approach is based on the "guilt-by-association" principle. It consists in grouping together genes (of known and/or unknown function) with a high expression correlation, with the hypothesis that this high correlation could be due to a common regulation, meaning that the genes play a role in the same biological process. We further searched for LNC hosting miRNAs genes, since the transcription of these LNC may result in the co-expression of their host miRNA[36] and highlighted interesting cases suggesting that the LNC and the miRNA could participate to the same biological process. Finally, starting from the premise that a gene expressed in one or a few tissues likely plays a role related to these tissue functions[37], we performed an in-depth analysis of LNC and PCG tissue specific expression. We showed that LNC are clearly more tissue-specific than PCG and pinpointed some specific features for the tissue-specific and divergent LNC:PCG pairs. The extended catalogue, in coordinates corresponding to the *Gallus_gallus-5.0* and the more recent *GRCg6a* assemblies, as well as all the information regarding the configurations of gene pairs and gene tissue-specificity are available in the Supplementary material of this article, and also on the FR-AgENCODE website (http://www.fragencode.org/). The files in this website will be regularly updated when new information or new version of the chicken genome assembly is available.

Results

Extension of ensembl gene catalogue with LNC gene models. The Ensembl gene catalogue (v94, December 2018) contained 24,881 genes (38,118 transcripts). Among these, 18,346 were annotated as protein-coding genes (PCG) and 4,643 as lncRNA (LNC) genes. Here we enriched this catalog by combining four complementary data sources: NCBI, ENCODE and ALDB LNC databases and the INRA catalogue newly generated in this study using 364 RNA-seq samples from 3 tissues. Using the pipeline "STAR – Cufflinks – Cuffmerge" and the "FEELnc" LNC prediction tool as described in Material & Methods section, we modelled 14,760 new genes composed of 1,199 PCG and 13,009 LNC genes. Among the LNC gene models, 7,265 were mono-exonics and 5,744 were multi-exonic. The expression of these new models was well above the background noise (Fig. 1a, "INRA" versus "Noise") for both LNC and PCG, and for both mono-exonic and multi-exonic models, corroborating their existence. As expected, LNC models were less expressed than PCG models. We then combined this set of new models (the *INRAGALG* models) with the ones from NONCODE[27], NCBI[26] and ALDB[28] public databases, to further extend the reference chicken Ensembl catalogue. The combination of these different catalogues is relevant since they are complementary as indicated by the low percentage of 1 bp-or-more overlapping transcripts between catalogues (from 3.2% to 29.1%) (Fig. 1b and Supplementary Figure S1 with more stringent overlapping criteria). The strategy used to add these four catalogues to the Ensembl catalogue is described in the Material and Methods. Briefly, we sequentially added the gene models from each database to a growing catalogue, keeping only the genes that had no same-strand 1 bp overlapping transcripts with a gene already present in the growing catalogue. The order of addition of the databases was determined using the CAGE peaks detected in 2017 by Lizio et al.[38] that corresponds to the transcription start site (TSS) of the transcripts: we calculated the percentage of LNC transcript models having their 5′ extremity within ± 30 bp of a CAGE peak, suggesting a good modelling, at least in the 5′-end. The corresponding percentages were 7.3%, 5.5%, 5.3% and 4.2% for INRA, ALDB, NCBI and NONCODE respectively, giving the order of addition of these sources. For the ± 100 bp criteria, these percentages were slightly higher (11%, 10%, 10% and 7% respectively). These values are low when compared to the 41% and 50% for the PCG transcripts using the ± 30 bp and ± 100 bp criteria respectively. Overall, these results are consistent with those of Derrien et al.[17] who found 15% for human LNC transcripts *versus* 55% for human PCG transcripts, using a ± 100 bp criteria. Lastly, we added to the catalogue, genes that we have recently produced in the multi-species FR-AgENCODE project[21,29]. The process leads to the generation of an extended catalogue of 52,075 chicken genes, with 30,084 LNC and 19,545 PCG genes, including all the Ensembl genes (4,643 LNC, 18,346 PCG and 2,446 other Ensembl genes, including the 1,705 small non-coding genes) (Fig. 1c). It gathers 6.5 times more LNC genes and 1.1 time more PCG, compared to Ensembl alone. This extended catalogue can be found in the form of a GTF file in Supplementary Data S1, and is available on the FR-AgENCODE project website[29]. The LNC density per chromosome was correlated with the density of PCG (Fig. 1d, Spearman $\rho = 0.90$, $p = 1.1 \times 10^{-06}$), with a higher density on the micro-chromosomes compared to the macro-chromosomes: 32 LNC and 36 PCG per Mb for the micro-chromosomes (chr 11 to 33) versus 22 and 12 per Mb respectively for the macro-chromosomes (chr 1 to 5).

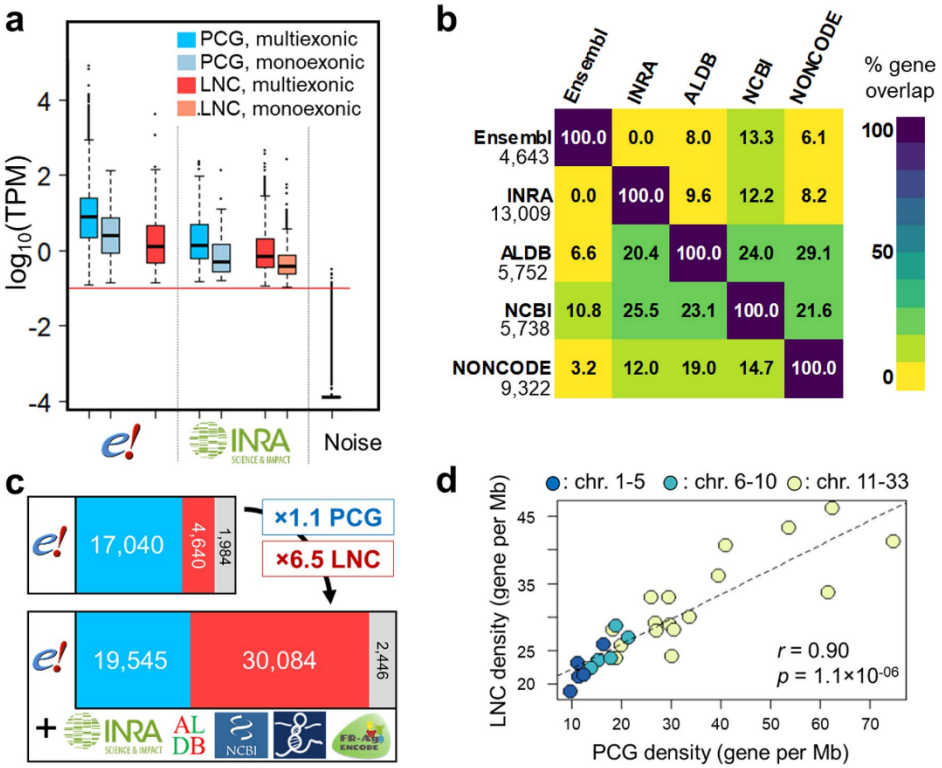


Figure 1. Extended gene catalogue features. (**a**) Expressions of the newly modelled genes compared to expression of the Ensembl genes and background noise, here given in the liver. The red line corresponds to the 0.1 TPM threshold. (**b**) Heatmap of the overlap between databases expressed in % of LNC (in line) shared among databases (in column), using 1 bp-or-more overlap. The number of LNC per database is mentioned in line. (**c**) The extended catalogue gathers 6.5 × more lncRNA genes and 1.1 × more PCG compared to Ensembl alone. (**d**) Correlation of LNC gene density to protein-coding gene density across the chicken macro-, medium- and micro-chromosomes. TPM: Transcript Per Million, chr: chromosome, Mb: Megabase.

Gene expression across chicken tissues and classification of the LNC with respect to the closest PCG. Using the extended catalogue, we quantified the expression of the LNC and PCG models in the 21 and 5 tissue datasets, (hereafter called "21 T" and "5 T", respectively). Out of the 52,075 genes of the extended annotation, on average 80% of the genes were expressed in at least one tissue of one dataset: 17,438 PCG (89%); 22,000 LNC (77%) and 777 other biotypes (32%). Interestingly, in the 21 T dataset similar numbers of expressed PCG were found across tissues (from 11,963 in muscle to 14,983 in ileum) while the number of expressed LNC are more variable (5,492 in kidney to 15,534 in duodenum) (Fig. 2a).

On average, LNC were 18 times less expressed than the PCG (Wilcoxon test, p-value $< 2.2 \times 10^{-16}$) (Fig. 2b), consistent with studies in human[17], dog[39] and chicken[22]. Among the 21,493 LNC and 17,241 PCG expressed in at least one tissue of the 21 T dataset, we observed 59% LNC (12,655) and 94%PCG (16,164) with expressions ≥ 0.5 TPM and 41% LNC (8,779) and 90% PCG (15,506) with expressions ≥ 1 TPM. The hierarchical clustering using the LNC expressions from the 21 T dataset shows biologically meaningful relationships among tissue types (Fig. 2c). The analysis grouped together tissues related to the nervous system (cerebellum and optical lobe), the immune system (spleen, thymus, bursa of Fabricius and the Harderian gland), the respiratory tract (trachea and lung, in green) and the digestive tract (pancreas, ileum, cecal tonsil and duodenum, in purple).

LNC models were classified with respect to the closest coding gene in the genome, according to the international gold-standard lncRNA classification provided by the GENCODE consortium[40] and using the FEELnc classification tool[41]. Results are summarized in Fig. 2d; the details gene-by-gene can be found in Supplementary Data S2. Of the 30,084 lncRNA genes of our extended catalogue, 23,129 genes were classified, the rest being genes either located alone on a contig, or without PCG in a 100 kb window (named "unclassified", see "Methods"). As expected, most LNC were intergenic (79%) and 21% were genic. Among the intergenics, the median distance between genes in the divergent LNC:PCG pairs was significantly lower compared to the distance within the convergent or same-strand pairs (3,957 bp *vs.* 5,130 bp, $p < 4.73 \times 10^{-15}$ and 3,957 bp *vs.* 6,005 bp, $p < 2.2 \times 10^{-16}$, Wilcoxon test). Splitting the intergenic genes in two classes based on distance[22] ("close": ≤ 5 kb and "distant": > 5 kb)

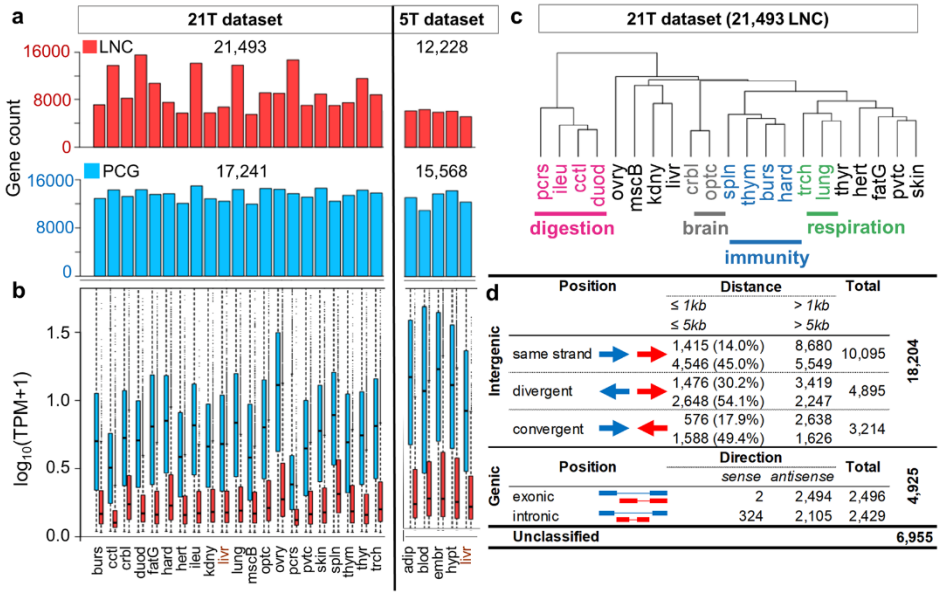


Figure 2. Overview of the extended catalogue in terms of expression and genomic configuration. (**a**) Number of LNC (red) and PCG (blue) expressed in each tissue: in the 21 T dataset, were expressed between 11,963 (muscle) and 14,983 (ileum) PCG with a median of 13,691 and between 5,492 (kidney) and 15,534 (duodenum) LNC with a median of 8794. In the 5 T dataset, were expressed between 11,024 (blood) and 14,313 (hypothalamus) PCG with a median of 13,165 and between 5127 (liver) and 6,319 (blood) LNC with a median of 6,043. (**b**) Boxplot of expression levels of LNC (red) and PCG (blue) in each tissue. (**c**) Hierarchical clustering of the 21 tissues from the 21 T dataset based on 21,493 LNC expressed in at least one tissue. Clustering performed using "1—Pearson correlation" distance and "ward" aggregation criteria. (**d**) Overview of the classification of LNC genes with respect to their closest PCG. For each configuration, the first line shows the numbers for a 1 kb threshold and the second line for a 5 kb threshold. Abbreviations in panels (**a**–**c**) stand for: burs: bursa of Fabricius, cctl: cecal tonsils, crbl: cerebellum, duod: duodenum, fatG adipose tissue around the gizzard, hard: harderial gland, hert: heart, ileu: ileum, kdny: kidney, livr: liver, lung: lung, mscB breast muscle, optc: optical lobe, ovry: ovary, pcrs: pancreas, pvtc: proventriculus, skin: skin, spln: spleen, thym: thymus, thyr: thyroid gland, trch: trachea.

shows an enrichment of close divergent compared to the convergent or same strand genes (54.1% versus 49.4% and 45%; $p < 3.9 \times 10^{-5}$ and $p < 2.2 \times 10^{-16}$ respectively, Fisher test). These two observations regarding the divergent pairs are supportive of widespread bidirectional transcription.

Co-expression differences of the LNC:PCG pairs according to their genomic configuration. In order to detect biologically meaningful relationships between LNC and PCG, we studied the expression correlations across tissues of the LNC:PCG pairs using the 21 T dataset which had the higher number of tissues. The results are displayed on Fig. 3a. We found 15,439 pairs expressed in at least one tissue among the 23,129 classified pairs. Out of these, 3,358 had a significant correlation in terms of expression (absolute value of Spearman $\rho \geq 0.55$, $p_{FDR} < 0.05$) with only 39 pairs having a negative correlation (Supplementary Table S1). We observed among the 3,358 co-expressed pairs a highly significant enrichment of divergent (14%), same-strand (32%) and sense intronics (49%) configurations, and a significant enrichment of antisense (9%) compared to convergent (7%). Furthermore, focusing on intergenic pairs, we observed a significant enrichment in co-expressed pairs separated by 5 kb or less compared to those separated by more than 5 kb for the divergent (16% versus 12%) and same-strand configurations (40% versus 25%) but not for the convergent configuration (8% versus 6%). Interestingly, comparing with co-expressed PCG:PCG pairs, we found 4,305 significantly co-expressed pairs, of which only 9 had a negative correlation. In these 4,305 pairs, we observed a significant enrichment in pairs separated by 5 kb or less versus more than 5 kb for all configurations, except convergent (Supplementary Table S2).

As an illustration, one example of positive correlation of LNC:PCG pairs is displayed in Fig. 3. The *PXDC1* (PX Domain Containing 1) gene shows a high expression correlation with the new INRAGALG00000007244 LNC model (Fig. 3c), which is in divergent configuration at 121 bp (Fig. 3b top). This pair is conserved in human and mouse (Fig. 3b, middle and bottom). In human, two close LNC genes are annotated in antisense and divergent configuration of *PXDC1*. Using the GTEx dataset[42], we found that both human LNC genes display a high expression correlation with the human *PXDC1* coding gene (Fig. 3d, top and bottom). Their expression pattern and their proximity suggest that they may represent one single gene and not two as annotated. The

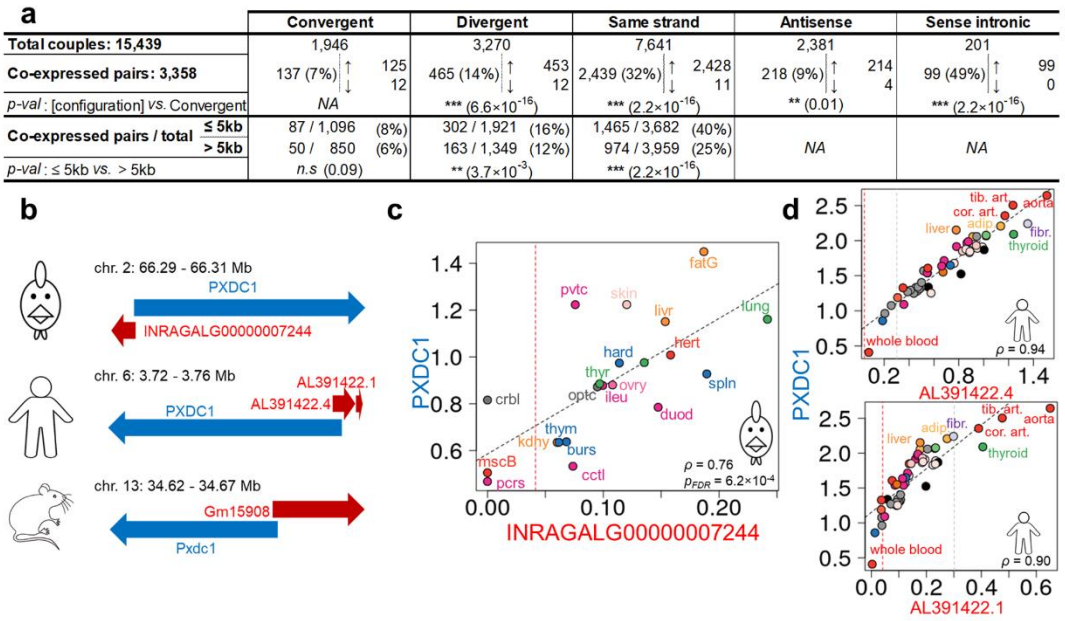


Figure 3. Classification and co-expression using the extended catalogue. (**a**) Overview of the correlations between the LNC and the PCG genes from all the expressed pairs, according to the different classes and the distance between the two genes of the pair. "*p-val: [configuration] vs. Convergent*" is the p-value of a Fisher test for the enrichment in significantly correlated pairs from the configuration in column *versus* the convergent configuration. "*p-val: ≤ 5 kb vs. > 5 kb*"is the p-value of a Fisher test for the enrichment of significantly correlated pairs at less than 5 kb *versus* more than 5 kb. (*: $p\text{-val} \leq 0.05$; **: $p\text{-val} \leq 0.01$; ***: $p\text{-val} \leq 0.001$; *n.s*: not significant). (**b**) Conservation of the genomic configuration of *PXDC1* and its closest LNC in chicken (top), human (middle) and mouse (bottom). In human, two close LNC are present. (**c**) Expression correlation in $\log_{10}(\text{TPM}+1)$ between INRAGALG00000007244 and PXDC1 across 21 tissues of the 21 T dataset. Tissues abbreviations are the same as in Fig. 2. (**d**) Conservation of the correlation between PXDC1 and its closest LNC in human in $\log_{10}(\text{TPM}+1)$. *NA* not applicable.

LNC:*PXDC1* pair represents a nice example of a conservation across species of both the genomic configuration and co-expression. The existence of the new chicken INRAGALG00000007244 LNC model was confirmed by RT-PCR through a clear amplification using lung RNA (Supplementary Figure S2), followed by sequencing.

LNC host small non-coding genes. We classified miRNAs and other small RNAs (Small nucleolar: snoRNAs, and small nuclear: snRNAs) genes with their closest LNC using FEELnc, in order to find LNC that host such genes. We found that 0.6% of the LNC (185 out of 30,084) hosted one or more miRNA, and that 16% of miRNA (177 out of 1,116) were hosted by one or more LNC. This is consistent with the literature in human in which 17.5% of the miRNA are located in LNC[36]. Similarly, for small RNAs, we identified 42 LNC (0.14% of total) hosting 58 small RNA genes, of which 48 snoRNAs (19% of snoRNAs) and 10 snRNAs (9% of snRNAs). These results are provided in Supplementary Data S2. Focusing on the 185 LNC hosting miRNA, we studied their co-expression with PCG across the 21 tissues of the 21 T dataset. Out of these 185 host LNC, 126 were expressed in at least one tissue of the 21 T dataset, and had between 0 and 939 correlated PCG among the whole dataset (spearman $|\rho| \geq 0.8$, $p_{FDR} \leq 0.01$), all positively. The selection process described in Material and Method left a list of 10 miRNAs cited in the literature and for which and 75% of the correlated PCG had a HGNC (Supplementary Tables S3–S13). Among these 10 host LNC, we found two cases in which the host LNC was conserved in human and mouse, and the correlated PCG were enriched in functions similar to those of the miRNA. The first example is reported in Fig. 4. The LNC INRAGALG00000001802 hosts the gga-mir-155, in a similar genomic configuration as MIR155HG and Mir155hg that host MIR155 and Mir155 in human and mouse, respectively (Fig. 4a). In chicken, we found that gga-miR-155 was highly expressed in the spleen and to a lesser extent in the thymus using the expression data from Chickspress[43] (Fig. 4b, top). In the 21 T dataset, INRAGALG00000001802 is mostly expressed in immunity-associated tissues (*i.e.* bursa of Fabricius, Harderian gland, spleen and thymus) and to a lesser extent in digestive system-associated tissues (caecal tonsil, duodenum, ileum and pancreas), as showed in Fig. 4b, middle. A similar pattern was observed in human for MIR155HG using the GTEx data, with a notable expression in the lymphocytes and the spleen (Fig. 4b, bottom). The 118 PCG highly correlated with INRAGALG00000001802 ($\rho \geq 0.8$) had as top5 enriched KEGG terms, terms associated with immunity (Fig. 4c). Figure 4d, top, provides some chicken LNC:PCG co-expression for four immunity-related genes taken

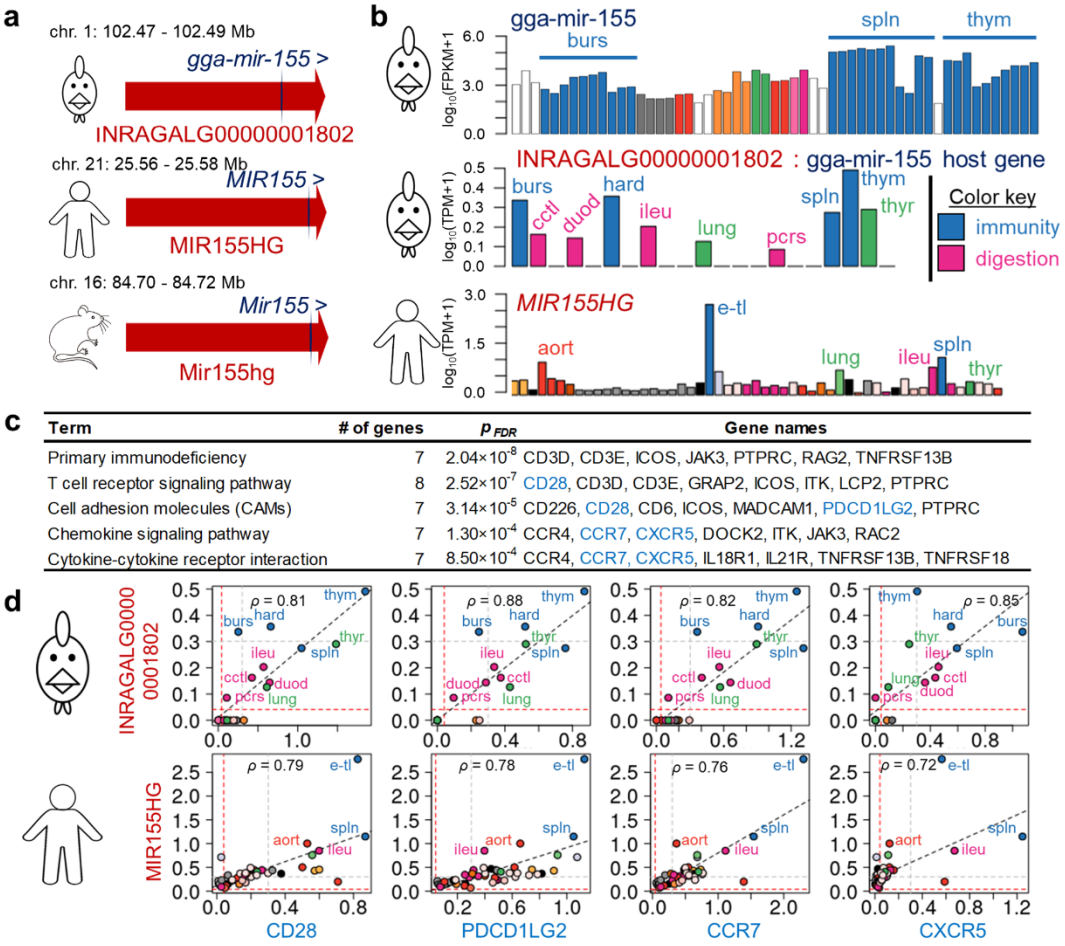


Figure 4. Conservation of genomic location and function of a miR host LNC across species. (**a**) INRAGALG00000001802 hosts gga-mir-155 and is conserved in human (MIR155HG) and mouse (Mir155hg). (**b**) Gga-mir-155 (top) and its host LNC, INRAGALG00000001802 (middle), are mostly expressed in immunity-related tissues in chicken, similarly to MIR155HG in human (bottom). Gga-mir-155 expression is expressed in $\log_{10}$(FPKM + 1) and the 55 Chickspress database tissues are ordered as the list available in Supplementary Table S22A, INRAGALG00000001802 and MIR155HG expressions are expressed in $\log_{10}$(TPM + 1) and the 53 GTEx project tissues are ordered as the list available in Supplementary Table S22B (**c**) Top 5 enriched KEGG terms supported by more than 5 genes associated to the PCG correlated to INRAGALG00000001802. PCG in blue are used in next panel. (**d**) Co-expression of four PCG from previous panel with INRAGALG00000001802 in chicken (top) or MIR155HG in human (bottom). Abbreviations in panels (**b**) and (**d**) stand for: aort: aorta, burs: bursa of Fabricius, cctl: cecal tonsils, duod: duodenum, e-tl: EBV-transformed lymphocytes, hard: harderial gland, ileu: ileum, lung: lung, pcrs: pancreas, spln: spleen, thym: thymus, thyr: thyroid gland.

as example. Interestingly, their human 1-to-1 orthologues were also well-correlated with the MIR155HG LNC, with the highest expression in the immunity-related tissues (Fig. 4d, bottom), suggesting that the mode of action responsible for this correlation in chicken is conserved in human. Interestingly, we found among the 118 PCG only two targets of the human hsa-miR-155-5p (*LAT2* and *ITK*), identified in both miRTarBase (with support type indicated as "weak") and mirDB, and one target of human hsa-miR-155-3p (*TXK*) using mirDB. *ITK* was also identified as a target of the chicken gga-miR-155 in mirDB (target prediction score = 98). The new chicken INRAGALG00000001802 LNC model was confirmed by RT-PCR through a clear amplification using spleen RNA samples (Supplementary Figure S2), followed by sequencing. The second case concerned gga-mir-124a-2, hosted in sense of intron by NONGGAG008930, which was expressed in the optical lope and cerebellum, and was correlated with 89 PCG enriched in genes associated to the development of the nervous system. In human, MIR124-3, of which gga-mir-124a-2 is a one-to-many ortholog with the same synteny, is hosted in antisense of

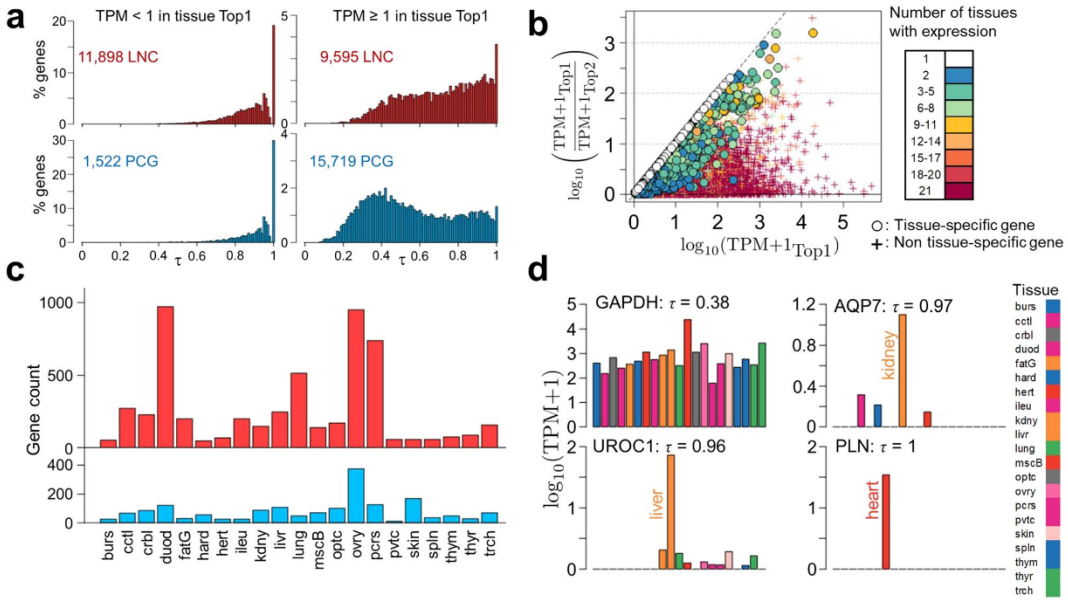


Figure 5. Overview of the tissue-specificity of PCG (in blue) and LNC (in red) in the 21 T dataset. (**a**) Distribution of τ values for LNC (top row) and PCG (bottom row) for two expression levels: TPM < 1 (left column) and TPM ≥ 1 (right column). The total number of corresponding genes is indicated in each plot. (**b**) For all the expressed genes (LNC and PCG), ratio of expression between the tissue with the highest expression ("Top1") and the second tissue with highest expression ("Top2") as a function of the expression in the tissue Top1. Colour indicates the number of tissues in which each gene is expressed, and the shapes differentiate the genes with a τ ≥ 0.95 (dots) versus τ < 0.95 (crosses). (**c**) Number of LNC (red) and PCG (blue) which are tissue-specific (τ ≥ 0.95) in each of the 21 tissues. (**d**) Expression profiles of 4 genes: the ubiquist *GAPDH* (top left), the tissues-specific *AQP7* and *UROC1* (top right and bottom left) and the heart-exclusive *PLN* (bottom right), expressed only in the heart. Tissues abbreviations are the same as in Fig. 2.

intron by a LNC expressed in brain parts and testis in the GTEx dataset. In mouse, *Gm27032* and Mir124a-3 are located on the same strand and have a small overlap. This LNC was well correlated with the human 1-to-1 ortholog of some of the 89 chicken PCG correlated with NONGGAG008930. Using Chickspress data, we found that gga-miR-124a, in particular the −3p transcript, was expressed in cerebellum, cerebrum and hypothalamus of adult chicken. Among these 89 PCG, we found only 3 targets of the chicken or human miRNAs. All these information are summarized in Supplementary Figure S3.

LNC are more tissue-specific than PCG. We studied tissue-specificity using the τ metrics[44], shown by Kryuchkova et al.[45], to have a good correlation between datasets, a good biological relevance and to be robust to normalization methods, compared to other metrics such as the Gini coefficient[46] or PEM[47]. The τ metrics associates to each gene a score ranging from 0 to 1. Zero corresponds to a gene that would be expressed at the same level in every tissue, while 1 corresponds to a gene that is expressed only in one tissue. Genes with a τ close to 1 usually show an important expression difference between the tissue with the highest expression and the other tissues in which it can be also expressed. It is noteworthy that the relevance of an analysis related to tissue specificity depends on the number of tissues studied. That is why we conducted this study using the 21 T dataset composed of a high number of tissues. Figure 5a shows the repartition of the τ values of the LNC (top row) and PCG (bottom row) for two expression levels: TPM < 1 in the tissue with highest expression (left) or TPM ≥ 1 in the tissue with highest expression (right). For the LNC, 55% of the genes had TPM < 1 and 45% had TPM ≥ 1, while these proportion were 9% and 91% for the PCG. For both the LNC and the PCG with TPM < 1, the distribution of the τ values is clearly right-leaning, showing only one hump close to 1 and a high number of genes with a τ equal to 1. With TPM ≥ 1, PCG and LNC show different τ-values distributions. For the PCG (Fig. 5a, bottom right), we clearly observe two peaks around τ = 0.4 and τ = 0.95, and a peak at τ = 1. The first hump corresponds to relatively ubiquitously expressed genes, with among them the well-known housekeeping gene Glyceraldehyde-3-Phosphate Dehydrogenase (*GAPDH*, τ = 0.38). The second hump corresponds to tissue-specific genes, with for example the Aquaporin 7 (*AQP7*, τ = 0.97) or the Urocanate Hydratase 1 (*UROC1*, τ = 0.96). Finally, the peak at τ = 1 corresponds to genes expressed in only one tissue (tissue-exclusive), such as the Phospholamban (*PLN*, τ = 1) expressed only in the heart, which is the principal regulator of the Ca2/-ATPase of cardiac sarcoplasmic reticulum[48]. The LNC (Fig. 5a, top right) presented a distribution of the τ values that is

clearly right-leaning, showing only one hump close to 1 and a high number of LNC with a τ equal to 1. Interestingly, the same patterns were observed for both LNC and PCG with expression data across 26 tissues from the dog species[39] (without consideration for the expression level, see Supplementary Figure S4). These distributions suggest a higher tissue-specificity for the LNC than the PCG: 6% of PCG with TPM ≥ 1 and 13.4% of LNC with TPM ≥ 1 had a τ value ≥ 0.95. We further characterized the genes having a tissue-specific expression, i.e. with a τ ≥ 0.95, a stringent threshold of tissue-specificity[39] by analysing at the same time their expression across tissues and the number of tissues in which they are expressed. The Fig. 5b shows for all genes the expression difference between the tissue with the highest expression (named "Top1") and the second tissue with highest expression (named "Top2") on the Y-axis, as a function of the expression in the Top1 on the X-axis. The number of tissues in which each gene is expressed is given by the colour of the dot. Most of the tissue-specific genes are expressed either in only one (n = 3,427 genes, white dots) or in a few tissues (2 to 5 tissues, n = 3,579 genes). The few genes that are expressed in 9 to 11 tissues (n = 62, yellow dots) or even 12 to 16 tissues (n = 13, orange/red dots) show a high level of expression in Top1 and are close to the diagonal, where Y = X, meaning that the difference between Top1 and Top2 is close to the value of Top1, *i.e.* their expression in Top2, and therefore in all the other tissues, is very weak (Supplementary Table S14). Overall, we found that 25.2% of the LNC (n = 5,422) and 9.8% of the PCG (n = 1,713) were tissue-specific in the 21 T dataset. Interestingly, the analysis of the 5 tissues in the 5 T dataset, revealed similar patterns for LNC and PCG, with 43.1% (n = 5,271) and 8.4% (n = 1,150) of tissue-specific genes, respectively (see Supplementary Figure S4). As expected, these percentages of tissue-specific genes are higher than what was observed in the 21 T dataset, due to the smaller number of tissues in the 5 T dataset. In each tissue, we found between 11 (proventriculus) and 375 (ovary) tissue-specific PCG with a median of 67 (Supplementary Table S14), and between 46 (harderian gland) and 972 (duodenum) tissue-specific LNC with a median of 156 (Fig. 5c, Supplementary Table S15). For each tissue, we realized a gene ontology (GO) terms enrichment analysis with the tissue-specific PCG (Supplementary Table S16). For example, the ovary-specific genes (tissue with the most tissue-specific PCG, see Fig. 5c) were enriched in 155 biological processes GO terms, of which "reproductive process" (GO:0,022,414, $p_{FDR} \leq 9.60 \times 10^{-15}$) or "sexual reproduction" (GO:0,019,953, $p_{FDR} \leq 5.75 \times 10^{-8}$). In the heart, the heart-specific genes (second tissue with the most specific PCG, see Fig. 5c) were enriched in 48 GO terms, of which "cardiac muscle tissue development" (GO:0,048,738, $p_{FDR} \leq 2.55 \times 10^{-9}$) or "blood circulation" (GO:0,008,015, $p_{FDR} \leq 1.7 \times 10^{-3}$). The expressions of three tissue-specific and one ubiquitously expressed are shown in Fig. 5d. *GAPDH* (most left) clearly shows a tissue-ubiquist expression pattern, with overall similar levels of expression across all tissues, albeit with a peak in muscle consistently with the literature[49]. *AQP* and *UROC1* have similar τ values, even though the former is expressed in four tissues while the latter is expressed in 10 tissues. The high difference in expression level between the Top1 and Top2 tissues of *UROC1* (liver and kidney, respectively) explains its tissue-specific classification. Finally, *PLN* clearly shows a tissue-specific pattern, being expressed in the heart only. More generally, for the tissue-specific genes (τ ≥ 0.95), the mean expression fold change between the Top1 and Top2 tissues is equal to 4.

We then investigated the liver-specific genes, this tissue being present in both 21 T and 5 T datasets. We found 247 (94 PCG; 145 LNC and 8 other biotypes) and 1,307 (326 PCG; 956 LNC and 25 other biotypes) liver-specific genes in the 21 T and 5 T datasets, respectively. We found that 208 genes (71 PCG and 130 LNC) were common to both datasets, meaning that the liver-specific genes from the 21 T dataset are almost included in those of the 5 T dataset. These genes are available in the Supplementary Table S17. Keeping only the genes with a unique HUGO identifier, the KEGG term enrichment analysis of the 54 common PCG revealed six KEGG terms ($p_{FDR} \leq 0.05$) provided in Supplementary Table S18, all of them related to liver function.

Tissue specific pairs in antisense or divergent configurations are enriched in DNA-binding transcription factors. Combining the tissue-specificity of both LNC (n = 5,422) and PCG (n = 1,713) from 21 T dataset with the classification of LNC:PCG pairs, we found 100 pairs for which both members were tissue-specific and had the same "Top1" tissue (Supplementary Table S19). The GO term analysis of these 100 genes (of which 45 PCG with a known identifier) revealed molecular functions related to "DNA-binding transcription factor activity, RNA polymerase II-specific" (GO: 0,000,981), supported by 14 genes: *ALX4*, *BARHL1*, *EMX1**, *HMX1**, *HOXD3**, *LHX5**, *MNX1*, *SOX1**, *TBX4**, *TBX5*, *TLX1*, *ZIC1*, *ZIC2* and *ZIC4* (Supplementary Table S20). Interestingly, these genes are enriched in the divergent configuration. In fact, in the 100 pairs list, only 12 genes belonged to the divergent class and six of them (indicated in the text with an asterisk) are related to the GO term abovementioned. Five out of the 6 LNC:PCG pairs seem to be conserved as antisense or divergent configurations in human and/or mouse as shown in Fig. 6b for *LHX5*, *HOXD3*, *TBX4* and in Supplementary Figure S5 for *EMX1* with the human ENSG00000278060 LNC, and *Hmx1* with the mouse ENSMUSG00000055944 LNC. Note that for *SOX1*, the closest LNC of human and mouse *SOX1* is an LNC-OT, i.e. an LNC in the same strand of the *SOX1* gene whereas in chicken both members of the LNC:*SOX1* pair are in opposite strands. Even if the LNC sequences are lowly conserved between distant species[50], we found six significant blast hits of length 41 to 253 in human and one sequence of length 296 bp in mouse similar to the one of chicken *SOX1* divergent LNC at 5380 bp and 4184pb, respectively, upstream of human and mouse *SOX1* gene on the opposite strand (Fig. 6b, bottom, Supplementary Table S21). Using the GTEx data[42], we analysed the co-expression *LHX5*, *HOXD3*,*TBX4* and SOX1 pairs available, the two other genes, *EMX1* and *Hmx1*, being absent in the GTEx dataset. Expression correlation of these four pairs in chicken and human, as well as their configurations are displayed in Fig. 6a and Fig. 6b respectively. The four chicken LNC models were confirmed by RT-PCR through a clear amplification using hypothalamus RNA samples for ENSGALG00000032696 and ENSGALG00000035091, kidney RNA sample for 107,053,814 and lung RNA sample for ENSGALG00000036551 (Supplementary Figure S2), and sequencing.

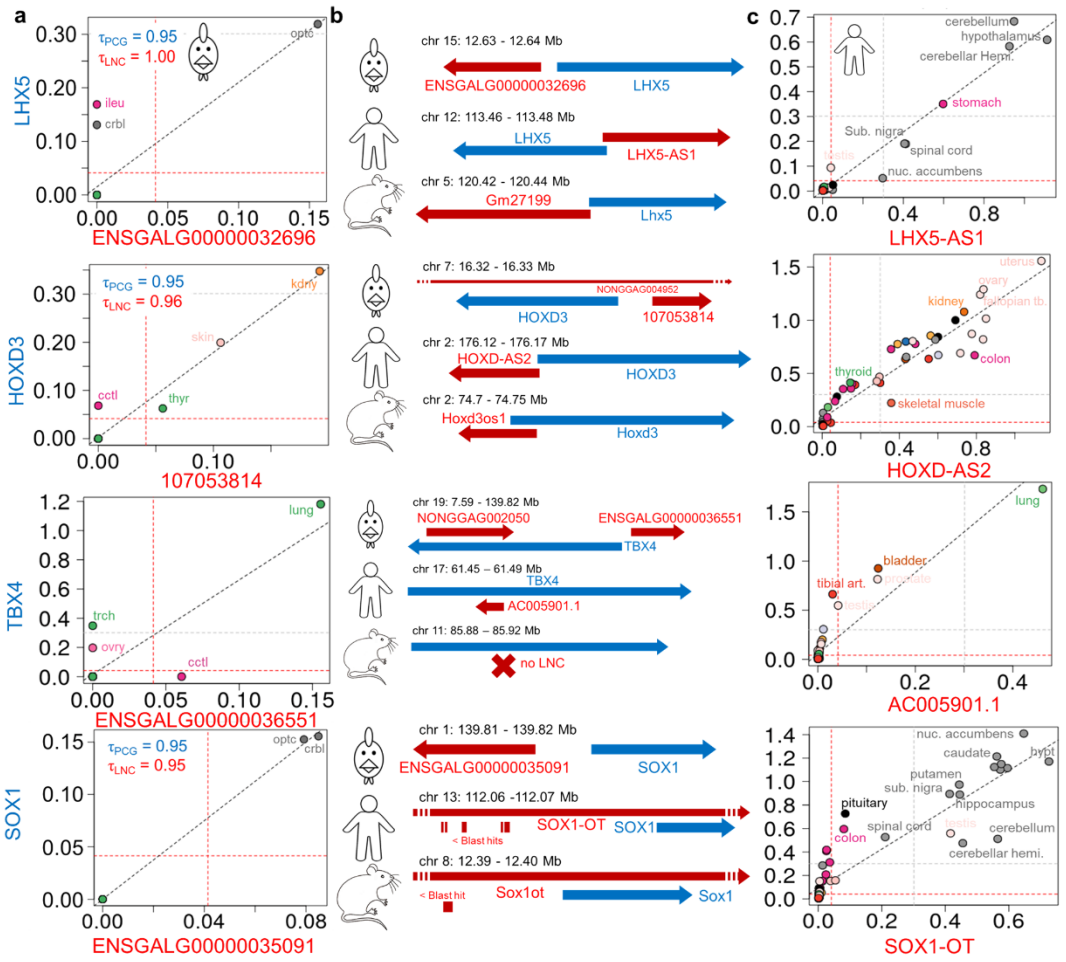


Figure 6. Tissue expression in chicken (**a**) and human (**c**) for 4 LNC:PCG pairs with similar genomic configurations between the two species (**b**). (**a**) Log_{10}(TPM + 1) expression of the LNC (X-axis) and the PCG (Y-axis) (top) for chicken across the 21 tissues of the 21 T dataset for the four chicken divergent pairs for which both members are tissue-specific in the same tissue. (**b**) Genomic configuration in three species of the LNC:PCG pairs. In red the LNC, in blue the PCG. For the LNC:SOX1 pair, were added the positions of hits (in minus strand) resulting in the mapping of ENSGALG00000035091 chicken LNC sequences to human and mouse genome on the opposite strand. (**c**) Expression in log_{10}(TPM + 1) of LNC:PCG pairs in 53 human tissues using the GTEx data. Tissues abbreviations are the same as in Fig. 2.

Conversion of the extended catalogue content to the GRCg6a newest version of the chicken genome assembly and the associated Ensembl v100 annotation. Since this work proposes a GTF build on the Ensembl v94 annotation in *Gallus_gallus-5.0* genome coordinates, we also generated an extended version of the Ensembl v100 annotation of the new *Gallus gallus* genome reference (*GRCg6a*), comprising genes present in our present catalogue that did not overlap genes from the Ensembl v100 annotation. The methods used for this update are described in details in the Material and Methods section. As a result, we added to the 24,356 genes of the Ensembl v100 annotation (16,878 PCG; 5,506 LNC; 1,972 others) 18,994 LNC gene models from the databases presented in this work, marked in the Supplementary Data S2. In addition, 3,179 models from the v94 annotation that were removed but that were mapped in the *GRCg6a* genome assembly. As a result, we generated an annotation comprising a grand total of 46,529 genes. Hence, users wishing to work on *GRCg6a*, can use this second GTF file available on the FR-AgENCODE website (http://www.fragencode.org/) to benefit from the Ensembl v100 enriched in LNC annotation.

Discussion

This work provides to the community an extended, LNC-enriched, gene catalogue of the chicken genome composed of 30,084 LNC, 19,545 PCG and 2,446 others genes, for a total of 52,075 genes. These data, in the form of a GTF file, are provided in Supplementary Data S1, and are ready to use by the community for RNA-seq expression analyses. We also provided a wealth of information in a table file (Supplementary Data S2 and its "read me" file, Supplementary Data S3). First, the configuration of the 30,084 LNC with respect to their closest PCG, plus the name of the PCG associated as well as the genomic distance between them. Second, for the 40,215 genes (of which 22,000 PCG and 17,438 LNC) expressed in one or both 21 T and 5 T datasets, we provide information relative to the tissue-specificity: the number of tissues in which the gene is expressed, the expression (in TPM) of the Top1 and Top2 tissues, the names of these tissues and the tissue-specificity values (τ values) in both datasets. For the LNC:PCG configurations, we also provide the spearman correlation of the expression for each gene pairs across the 21 and the 5 tissues of both datasets. All this information can be precious for researcher interested in one or several LNC to infer hypotheses about their expression and function. This GTF file, build on Ensembl v94 annotation, in Gallus_gallus-5.0 coordinates, and all related information are available on the FR-AgENCODE website (http://www.fragencode.org/), and they will regularly be updated, in particular when new assemblies are published. As a first update, we propose in this website an extended version of the Ensembl v100 annotation of the new *Gallus gallus* genome reference (GRCg6a), comprising genes present in this catalogue that did not overlap genes from the Ensembl v100 annotation, in GRCg6a coordinates.

We then performed different types of analysis in order to provide some functional annotation to these LNC. The analysis of the LNC:PCG configurations revealed different types of pairs, in proportions that were consistent with the literature in human[17] and in farm species (pig, cattle and chicken)[51], with a majority of intergenic genes (approximately 80%) of which 56% are in same-strand, the rest being in divergent and convergent configurations. Pairs in same-strand configuration should be considered with caution, since it is possible that the LNC is in fact a part of a not yet properly modelled PCG, especially when the gene couple is significantly co-expressed across tissues (as we show for 32% of the same strand pairs). Indeed, in non-model species, PCG isoforms are still poorly annotated. Only 38,118 isoforms are described in the chicken Ensembl v94 annotation for 24,881 genes in total, while more than 200,000 transcripts are modelled for 57,720 genes in total in the human Ensembl v94 catalogue. As an example, we recently showed that an LNC located at 978 bp in same-strand of *SCD1* was in fact part of this gene[52]. However, this situation is also true for close, same-strand LNC, as exemplified in Fig. 3d, in which the high correlation of the two human LNC with the PCG genes suggests that they in fact constitute one unique gene.

The divergent or genic configuration for LNC:PCG pairs associated to their co-expression may allow to formulate hypotheses on the function of LNC. Indeed, a significant expression correlation between two genomically close genes is an argument for the existence of a common regulation[53] or even a regulation of the PCG by the LNC[35]. In both cases, we can hypothesize that the LNC is involved in the same biological processes as the PCG[54,55]. These hypotheses inferred by co-expression combined to gene configurations should allow to better orientate molecular biology experiments in order to better understand the biological role of the LNC of even of the LNC:PCG pair. Interestingly, we found a significant enrichment of co-expressed pairs among the divergent LNC:PCG configuration (14% of these pairs) and among the sense of introns configurations (49% of these pairs) compared to the convergent configuration (7% of these pairs). This enrichment is more limited but remains significant for the antisense configuration with 9% of these co-expressed pairs, compared to the convergent configuration. Divergent pairs suggests the presence of a bidirectional promoter that regulates both genes[56], or even an effect of the LNC on the PCG expression[31,32], even though the precise mechanisms remain unclear. We highlighted an example with the divergent LNC:*PXDC1* pair, conserved between chicken and mammals. *PXDC1* is a gene for which the function is still poorly known. It was found to be repressed in the liver of mice and rats exposed to a pollutant of the dioxin class (TCDD, or 2,3,7,8-tetrachlorodibenzo-p-dioxin)[57]. Furthermore, it was shown that transcriptional activation of *PXDC1* using a CRISPR activation system increased the survival of an acute myeloid leukaemia cell line exposed to cytarabine, a molecule used in standard chemotherapeutic mix for this leukaemia, hence increasing the resistance of the cells to the chemotherapy[58]. The high expression correlation between *PXDC1* and its divergent LNC suggests that the latter plays a role in the same biological processes as *PXDC1*. Antisense LNC, for their part, have also been shown to regulate their host gene expression, at both the transcriptional level, by the inducement of chromatin remodeling[33], or at the post-transcriptional level[34]. Finally, sense intronic can arise from the splicing of a PCG or be produced from independent transcriptional unit[59]. They may regulate the expression of their host PCG, which are also associated to transcription regulation[60], as shown by Guil et al., who observed a down-expression of the gene *SMYD3* with the over-expression of a LNC from its intron[61]. In this study, we described 214 and 99 significantly co-expressed LNC:PCG pairs in antisense and sense intronic configuration respectively.

Interestingly, across all the LNC:PCG configurations, we found only a small number of pairs for which the correlation was significantly negative with 1.2% (39 pairs) out of the 3,355 significant correlations (with 0 for sense intronics and 12 for both the divergent and convergent configurations). Such a result was also observed in human[17], for which 0.11% of the pairs LNC:PCG showed correlation lower than − 0.5, and in dog[39], for which this percentage was 0.71%. Taken together, these results shared among different species suggest that LNC tends to act as positive, rather than negative, regulators or cofactors of the transcription of their closest gene, although there are well-known examples, such as *HOTAIR*[62] or *XIST*[63], that show that LNC can induce gene silencing. Interestingly, this trend was also found in PCG:PCG couples, for which only 0.15% (6 pairs) out of the 4,125 significant correlations were negative (3 divergent and 3 same-strand).

We found in chicken the same proportion of miRNA hosted by LNC than in human, 16% compared to 17.5%[36]. LNC hosting miRNA are thought to act as pri-miRNA (coined lnc-pri-miRNA in Dhir et al.[36]), which are the precursors of pre-miRNAs, themselves precursors of miRNAs[64]. Among the 10 LNC hosting miRNA

chosen here to be deeply analysed because of their functional annotation in the literature, we found two cases in which the LNC hosting a miRNA might be associated to the same biological process as the miRNA. Indeed, we observed that mir-155, which is involved in the normal immune function[65] and also expressed in immune-related tissues in chicken is hosted by a LNC that we also found to be highly expressed in the immune-related tissues and correlated in expression with multiple immunity-related genes in both chicken and human. Such results are consistent with a recent work conducted by Maarouf et al.[66] who showed in a human cell line that its host LNC, MIR155HG participate to the immune response to the influenza A virus (IAV). What is more, Maarouf et al.[66] showed that MIR155HG deleted of the sequence of MIR155 still significantly suppressed the IAV replication, clearly indicating that this LNC had an action on its own, not only by harbouring MIR155 sequence. While the immunity-related function of miR-155 and now of the LNC hosting this miR are known, the relationships between these two genes is not clear. The low number of miR-155 targets detected among the correlated genes tends to suggest that mir-155 does not directly act on the immunity-related PCG co-expressed with its LNC. The second example concerned gga-mir-124a-2, the human ortholog of which is MIR124-3, which is involved in neurogenesis[67,68]. We observed that gga-mir-124a-2 was expressed in different brain-parts in adult chicken, and found that its LNC hosting gene in chicken and human was also expressed in the brain-related tissues. In both species, the LNC were also correlated in expression with genes involved in the nervous system. Supplementary experiments are needed to better understand mechanisms underlying such PCGs and miR155 host LNC or of miR124 host LNC co-expression across tissues.

Our analysis of the tissue-specificity of the gene expression showed that LNC are more tissue-specific compared to PCG: 25.2% of LNC versus 9.8% of PCG for the dataset with 21 tissues, using the τ metrics with a threshold set at 0.95. The tissue-specificity of genes depends on multiple factors, such as the number of tissues analysed or the tissue-specificity metrics used. Nevertheless, the higher tissue-specificity of LNC compared to PCG is consistent with previous works conducted on several species, even if the ratio of tissue-specific genes in each biotype is variable. In human, Derrien et al.[17], using 16 tissues and counting the number of tissues in which each gene was expressed, found that approximately 27% of the LNC and 4% of the PCG were expressed in one or two tissues, while Cabili et al.[30], using 24 tissues and cell types, found more extreme values with 78% of LNC and 19% of PCG found to be tissue-specific using an entropy-based measure. Melé et al.[69], using the GTEx data composed of 43 healthy tissues and 13 brain subregions of cells lines also commented on the fact that PCG are generally ubiquitous while LNC are more typically tissue-specific. In dog, Le Béguec et al.[39], using 26 tissues and the τ metrics with a threshold set at 0.95, found that 44% of LNC and 17% of PCG were tissue-specific. It is noteworthy that the absence of the testis in our analysis may be responsible for an under-estimation of the total number of tissue-specific genes, since this tissue is well known for having a specific expression. Indeed, Melé et al.[69] found that ~95% of the approximately 200 genes expressed in only one tissue were expressed in the testis.

Concerning the liver, which is the only tissue common to both independent 21 T and 5 T datasets analysed in this work, the high intersection of liver-specific genes detected between both datasets shows a good reproducibility of the results. This result shows the robustness of the gene tissue-specificity despite the technical and biological variations between the two datasets, from RNA extraction to sequencing, and from lines to ages or breeding conditions. These common genes were in enriched liver functions as "Complement and coagulation cascades"[70], "Metabolic pathways", "Histidine metabolism"[71], "Tryptophan metabolism"[72], "Steroid hormone biosynthesis"[73] and "PPAR signalling pathway"[74].

Interestingly, we observed an enrichment of GO terms related to the regulation of transcription among the LNC:PCG pairs for which both members were tissue-specific in the same tissue, with an over-representation of the divergent pairs. Five LNC:PCG pairs for which the PCG codes for a transcription factor involved in embryonic development or cell lineage seemed to be conserved between chicken and mammals. First, they share the same divergent or antisense configuration between chicken and human and/or mouse, at the exception of SOX1 for which a LNC was found in divergent configuration in chicken, versus in sense of intron overlapping in human and mouse (called *SOX1-OT*), but for which blast hits were detected in divergent configuration, suggesting the existence of another LNC. Hence, it is difficult to conclude on the conservation of ENSGALG00000035091 even if a high co-expression and similar tissue-specificity between the human *SOX1* and *SOX1-OT* genes was observed. Second, the tissue specificities observed in chicken are consistent with the one observed in human using GTEx data for *LXH5* and *SOX1* as brain-specific, *TBX4* as lung-specific and *EMX1* as kidney-specific. *HMX1* is found to be skin-specific in chicken while it is testis and brain specific in human. Interestingly, we found *HOXD3* to be kidney-specific in chicken while in human its "top tissues" are part of the female reproductive tract, just before the kidney (Fig. 6c,). This tends to indicate an expression conservation within the kidney across species, but not in the reproductive tract between these two species diverging on reproductive functions (oviparity versus viviparity). Third, these common tissue specificities are consistent with the function of the proteins associated to the PCG of each couple. *LXH5* and *SOX1* have been shown to be involved in neural determination or differentiation[75,76]. In particular, *SOX1* is known to control the development of the neural ectoderm from which the optical lobe and cerebellum are derived in adults. *HOXD3* belongs the HOXD cluster known to regulate the metanephric kidney development in mammals[77,78]. Recently, *EMX1* was reported as a novel nephron segment regulator during embryonic kidney development[79]. Finally, different studies report an important role of *TBX4* in lung development[80–82].These results raises the question of the precise role of the LNC on the PCG in each of these tissue-specific pairs in their respective tissues.

Overall, the ability to obtain a comprehensive catalogue of the LNC in a species depends mainly on the number of tissues available, as shown in this work and in the literature[17,30,39] on developmental stages and in a lesser, but not negligible, extent on the experimental conditions, or age of the organisms studied. Another limiting factor in such endeavour is the cellular heterogeneity of the tissues. Indeed, tissues are composed of different cell types, and most of them likely contain blood cells that irrigate them. Single-cell RNA sequencing (scRNA-seq) could to be a method of choice to discriminate between cell types and therefore minimize heterogeneity bias.

However, scRNA-seq is currently limited in the number of detected transcripts, estimated to only 10–20% of the mRNA molecules actually present[83], and seems to perform poorly in the detection of low expressed genes, which is the case of LNC. These technical limitations will therefore have to be alleviated for scRNA-seq to be used in LNC detection and modelling.

As a conclusion, this study aims at providing an improved gene annotation enriched in LNC for the chicken species, information on their genomic configuration with respect to PCG and miRNA, and transcription patterns across tissues for the both PCG and LNC of the annotation. Such a catalogue and information associated will be useful to the community to work on LNC, for example, to unveil the molecular chain of causality that links variants located outside coding regions and phenotypes of interest. As an example, Plassais et al.[84] recently demonstrated the causative role of a point mutation in the exon of a LNC, located in divergent configuration with a PCG involved in neural development, in a form of neuropathy. Both genes appeared to be co-expressed, and the point mutation seemed to affect the binding of regulatory elements, leading to the reduction of their expression.

Methods

Biological samples used. For the gene modelling (i.e. the computational prediction of transcript structures in the genome), we used 364 RNA-seq samples from three chicken tissues (blood, liver and adipose tissue) with different physiological stages of broilers and layers, for which the raw RNA-seq data are available under the SRP079637 and PRJEB28745, PRJEB34310 and PRJEB34341.

For the expression study, we used 5 tissues (blood, adipose tissue, liver, hypothalamus and embryos) from a Rhode Island Red line with a minimum of 24 biological replicates per tissue (data available under the PRJEB28745). Such a dataset with an average of 80 M reads per biological replicate is referred to as the "5 T dataset" in the Results section. We also used 168 RNA-seq samples from the PRJEB12891, used by Ensembl for annotating the chicken *Gallus_gallus_5* reference genome, version v87 (December 2016) to v94 (October 2018). This dataset was composed of 21 tissues with 8 replicates per tissue with, on average, 15 M of reads per replicate representing one age (16–17 weeks old), one sex (female) of the same J-line strain. It is referred to as the "21 T dataset" in the "Results" section.

Ethics. All the animal experiments used in the present study were approved by the local ethical committee in animal experimentation of Val de Loire, France (authorization to experiment on living animals n°7740, 30/03/2012) and by the French Ministries of Higher Education and Scientific Research, and of Agriculture and Fisheries (Approval number: 2873–2,015,112,512,076,871). Animal experiments were conducted at the experimental farm PEAT under license number C37-175–1 for animal experimentation, in compliance with the European Union Legislation.

RNA collection and sequencing. RNA extraction and RNA sequencing were performed as described in Muret et al.[22] and Jehl et al.[85]. Briefly, tissues were sampled immediately before (blood) or after slaughter (adipose tissue, liver) and stored appropriately. RNAs were extracted following the kits or reagents manufacturer's instructions and stored at −80 °C. Total RNA was quantified using a NanoDrop ND-1000 spectrophotometer (Thermo Scientific, Illkirch, France). The A260/280 and A260/230 ratios were greater than 1.7 in all samples ensuring the RNA purity. The RNA quality was controlled using an Agilent 2100 bioanalyzer (Agilent Technologies France, Massy, France). The RNA integrity numbers were ≥ 7 for the adipose and the whole blood tissue, ≥ 8 for the hypothalamus, ≥ 8.5 for the liver and embryos. The sequencing was conducted in a stranded, paired-end 2 × 150 bp reads, on a HiSeq2000 or HiSeq3000 (Illumina).

INRAGALG gene modelling. RNA-seq reads were trimmed using cutadapt version 1.8.3. Reads were then mapped on the Ensembl *Gallus_gallus_5* reference genome using STAR[86] v.2.5.2b, following the multi-sample 2-pass mapping procedure[87], with the Ensembl v92 GTF file as input file for the generation of genome indices as described in Muret et al.[22]. The new transcript models were constructed as described in Foissac et al.[21]. Briefly, after read mapping and CIGAR-based softclip removal, each sample alignment file (BAM file) was processed with Cufflinks 2.2.1 with the max-intron-length (25,000) and overlap-radius (5) options, guided by the reference gene v92 annotation (–GTF-guide option). All cufflinks models were then merged into a single gene annotation using Cuffmerge 2.2. with the –ref-gtf option. Using the 364 RNA-seq samples, we modelled 25,085 putative new genes in addition to the Ensembl genes. To ensure reliability of the models, loci were selected based on their expression and their reproducibility across samples. First, 21,520 genes were selected with an expression greater than or equal to 0.1 TPM (Transcripts Per Million), a common threshold when working on lncRNA genes[88,89], known to be lowly expressed. However, we observed that some models exceeding this threshold were supported by one or two reads at most, whichever the replicate. Hence, in order to discard such models, we applied a more stringent criterion, consisting in keeping only the models supported by a least five reads in the samples of a given tissue, similarly to de Goede et al.[89]. These selection steps resulted in a dataset of 14,760 models, hereafter called *INRAGALG*.

LNC prediction. The discrimination between coding and long non-coding genes was realized using the FEELnc codpot module of the FEELnc (FlExible Extraction of Long noncoding RNAs, v0.1.0) tool[41,90], as described in Muret et al.[22]. Briefly, FEELnc codpot module calculates a coding potential score (CPS) for the assembled transcripts based on several predictors (such as multi *k*-mer frequencies and Open Reading Frame coverage) incorporated into a random forest algorithm[91]. In order to increase the robustness of the final set of novel lncRNAs and mRNAs, the option –spethres = 0.98 was set. The FEELnc model was trained with the chicken PCG and LNC transcripts from the chicken Ensembl v92 annotation.

Background noise evaluation. The background noise corresponds to the expression of a set of artificial loci randomly distributed across chicken chromosomes 1 to 33 using the bedtools shuffle function from the BEDTools suite[92]. These artificial loci had the same length distribution as the LNC genes and were positioned at least 5 kb out of the closest known transcribed regions.

External source pre-treatment and aggregation. *Ensembl* source: the *Gallus_gallus_5* Ensembl v92 reference annotation was downloaded from the Ensembl FTP website[106] as a GTF file. *INRA* source: the INRA GTF file was obtained and filtered as described in the two previous sections. *ALDB* source: the ALDB[28] v1.0 database containing 5,752 LNC genes was downloaded from the FANTOM project Zenbu viewer[93] page, as a BED file in *Gallus_gallus_5* version. *NONCODE* source: the NONCODE v5.0 database containing 9,322 LNC genes was downloaded from the appropriate website[94], the form of a BED file being in *Galgal4* version. The file was therefore translated into *Gallus_gallus_5* version using the LiftOver tool from UCSC[95], and the chromosome names were converted into Ensembl chromosome names. *NCBI* source: the NCBI database v103 containing 5,738 LNC genes was downloaded from NCBI FTP website[96] in the form of a GFF3 file in *Gallus_gallus_5* version. Chromosome names were converted into Ensembl chromosome names. Finally, the GFF3 file was converted into a GTF file as well as the BED files from ALDB and NONCODE sources. *FR-AgENCODE* source: the FR-AgENCODE annotation containing 6,089 LNC genes was obtained from Foissac et al.[21].

Aggregation of the sources: we sequentially added the gene models from each database to a growing catalogue, keeping only the gene models that had no overlap with a gene already present in the growing catalogue. Two gene models were considered as overlapping if one or more of their transcripts were on the same-strand and had one or more nucleotides in common. The gene model overlap between sources was assessed using the bedtools intersect function from the BEDTools suite[92]. The different sources were aggregated to the Ensembl annotation in the order presented in Results, which was determined by calculating the percentage of LNC transcript models having their 5′ extremity within a CAGE peak[38], suggesting that these transcripts were properly modelled, at least in their 5′-end. Note here that this sequential strategy was chosen since we observed that total aggregation of the overlapping models using the cuffmerge tool tended to create chimeric models composed of one or more known PCG Ensembl genes by dubious LNC models from another database.

LNC and miRNA classification using FEELnc. Long non-coding transcripts were classified relatively to the closest protein-coding transcript using FEELnc classifier tool[41,90] with default settings (maximal window of 100 kb). Briefly, using (*i*) the LNC transcript models from the extended annotation and (*ii*) the PCG transcript models, the tool uses a 100 kb sliding window and classifies each LNC transcript using its location and orientation relative to the closest PCG transcript. The results distinguish LNC types (whether genic or intergenic), then subtypes (overlapping, containing and nested subtypes for the genic type, and divergent, convergent and same strand subtypes for the intergenic type) and finally locations (exonic, intronic or upstream, downstream). If no PCG were found within the sliding window, the LNC is considered as "unclassified". From this transcript level classification, we generated a gene level classification. Generally, the different transcripts of a given LNC gene have the same classification relative to PCG. However, for the rare cases (4%) in which transcripts of a given LNC gene have different classifications relative to one or more PCG, we indicated these conflicts. The FEELnc classifier module was also used to classify the miRNAs present in the Ensembl annotation with respect to their closest LNC for identifying LNC hosting one or several miRNAs.

Gene expression quantification. FASTQ files were mapped on the Ensembl *Gallus_gallus_5* reference genome using STAR[86] v.2.5.2b, following the multi-sample 2-pass mapping procedure, with the extended GTF file as input file for the generation of genome indexes step as described in Muret et al.[22]. Samples were analysed by tissue. FASTQ files were previously trimmed for Illumina adapter using TrimGalore version 0.4.5. Expression was quantified with RSEM[97] v.1.3.0, using the extended GTF file at the gene-level[21,22,39]. This workflow is part of a snakemake[98] pipeline, available at Ref.[99]. Each gene was considered as expressed in at least one tissue of 5 T or 21 T dataset using the criteria TPM ≥ 0.1.

For the 21 T dataset, the 21 tissues and their four letters abbreviations are: bursa of Fabricius (burs), cecal tonsils (cctl), cerebellum (crbl), duodenum (duod), adipose tissue around the gizzard (fatG), harderian gland (hard), heart (hert), ileum (ileu), kidney (kdny), liver (livr), lung (lung), breast muscle (mscB), optical lobe (optc), ovary (ovry), pancreas (pcrs), proventriculus (pvtc), skin (skin), spleen (spln), thymus (thym), thyroid gland (thyr) and trachea (trch). For the 5 T dataset, the 5 tissues, blood, adipose tissue, liver, hypothalamus and embryos, were abbreviated as blod, adip, livr, hypt and embr respectively.

GTEx data analysis. The version 7 RNA-seq TPM data was downloaded from the GTEx website (https://gtexportal.org/home/). Each gene expression was normalized as $\log_{10}(\text{TPM}+1)$, and the mean for each of the 53 tissue was calculated. The list of the 53 tissues is available in Supplementary Table S22.

Tissue-specificity analysis. Tissue-specificity was assessed using the tau (τ) metric[44], which ranges from 0 (gene expressed at the same level in all tissues) to 1 (gene expressed in exactly one tissue), using the following formula: let $x_{g,t}$ be the expression of the gene g, in the tissue t, among T tissues. The τ value associated to a gene g is calculated using the following equation:

$$\tau_g = \frac{\sum_{t=1}^{T}\left(1-\hat{x}_{g,t}\right)}{T-1}, \quad \text{where} \quad \hat{x}_{g,t} = \frac{x_{g,t}}{\underset{1\le t\le T}{max}\left(x_{g,t}\right)}$$

with $x_{g,t}$ being the expression of the gene g, in the tissue t, among T tissues.

A gene was considered as tissue-specific for $\tau \geq 0.95$, as done in previous studies[39], and corresponding to a ratio of 4 between the first and the second tissues with the highest expressions.

Co-expression analysis. For each LNC:PCG pairs detected, we computed the Spearman correlation (ρ) between the expression values across tissues. *P*-values were corrected for multiple testing using the Benjamini–Hochberg method[100]. The false discovery rate was set to 0.05, corresponding to an absolute ρ value of 0.55 using the 21 tissues of the PRJEB12891 dataset, referred to as the "21 T dataset" in the Results section.

GO-terms and KEGG terms analysis. The enrichment analysis of Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) terms in each set of genes of interest was performed using the STRING tool[101] (https://string-db.org). Only the 1-to-1 human orthologous genes with a standardized HGNC name were submitted for the analysis. To investigate biological functions, we favoured KEGG terms when available, GO terms related to "Biological Process" otherwise. If no such enriched terms were found, we focused on the molecular functions of the genes using the GO "Molecular Function" terms.

LNC hosting miRNA analysis. For each expressed LNC containing one or more miRNA gene, we calculated the expression correlation between the LNC and all the expressed PCG. The correlation threshold was set at $|\rho| \geq 0.8$, corresponding to a false discovery rate inferior or equal to 0.05 with the 21 tissues of the PRJEB12891 dataset, referred to as the "21 T dataset" in the Results section. MiRNA expression data were obtained from the Chickspress website[43] in which 55 tissues and conditions are available, the list of which is provided in Supplementary Table S22. We further selected the LNC for which at least 75% of the correlated PCG had a HGNC and for which the hosted miRNA gene(s) were cited in the literature following a PubMed request of their name. Targets of the miRNAs were screened for in miRTarBase[102] and mirDB[103]. For the latter, we only considered targets with a target prediction scores ≥ 80, which indicates a high confidence of the prediction.

Biological validation by RT-PCR and sequencing. The existence of different LNC models was assessed by RT-PCR and sequencing using RNA extracted from different chicken tissues. Each LNC RNA was validated using the tissue in which it was the most expressed. The PCR primers and hybridization temperature used are indicated in Supplementary Table S23 for each analysed LNC. Reverse transcription (RT) was carried out using the high-capacity cDNA archive kit (ThermoFisher Scientific, catalog number: 4368814) according to the manufacturer's protocol. Briefly, reaction mixture containing 2µL of 10 × RT buffer, 0.8µL of 25 × dNTPs, 2µL of 10 × random primers, 1µL of MultiScribe Reverse Transcriptase (50 U/µL), and total RNA (2 µg) was incubated for 10 min at 25 °C followed by 2 h at 37 °C and 5 min at 85 °C. Dilution RT reaction was further used for PCR. 5 µl of cDNA samples were mixed with 5 µl of GoTaq Flexi Buffer 5 ×, 2µL of $MgCl_2$ solution (25 mM), 0.125 µl of GoTaq DNA Polymerase (5u/µl) (Promega, catalog number: M891), 0.5 µL of dNTPs 10 mM, 12.5 µl H20 and 1.25 µl of specific reverse and forward primers at 10 µM. Reaction mixtures were then incubated in a T100 Thermal cycler (Bio-Rad, Marne la Coquette, France). The amplification products are then deposited on a 2% agarose gel and sent for sequencing (Genoscreen) to verify their location in the chicken genome. LNC were considered as validated if a PCR amplification was observed and their location in the genome was the expected one.

Update of the catalogue in GRCg6a reference genome using the associated Ensembl v100 annotation. Transcripts sequences were extracted from the Gallus_gallus-5.0 FASTA file using gffread[104] v0.11.0 and mapped to the GRCg6a assembly sequence using GMAP[105] 2015-11-20 with default parameters. If none of a gene's transcripts had a unique position (mapped on different chromosomes or on different strands), the gene and its transcripts were removed from the annotation. The genes from our catalogue in Gallus_gallus-5.0 coordinates for which even one transcript overlapped an Ensembl gene from the v100 annotation were removed. This insures that the annotation that we provide in GRCg6a positions is composed of the full Ensembl v100 annotation, enriched in LNC gene models from the present catalogue. This GTF generated in coordinates GRCg6a, as well as the GTF in Gallus_gallus-5.0 coordinates can be found the FR-AgENCODE website (http://www.fragencode.org/overview.html).

Data availability

The raw transcriptomic data used in this work are available on ENA under accession numbers: PRJEB28745, PRJEB34310 and PRJEB34341, and on SRA under Accession No. SRP079637.

Received: 23 January 2020; Accepted: 11 November 2020
Published online: 24 November 2020

References

1. Brannan, C. I., Dees, E. C., Ingram, R. S. & Tilghman, S. M. The product of the H19 gene may function as an RNA. *Mol. Cell. Biol.* **10**, 28–36 (1990).
2. Brockdorff, N. *et al.* The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell* **71**, 515–526 (1992).

3. Sun, M., Gadad, S. S., Kim, D.-S. & Kraus, W. L. Discovery, annotation, and functional analysis of long noncoding RNAs controlling cell-cycle gene expression and proliferation in breast cancer cells. *Mol. Cell* **59**, 698–711 (2015).
4. Ng, S.-Y., Johnson, R. & Stanton, L. W. Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors: LncRNAs involved in neuronal differentiation. *EMBO J.* **31**, 522–533 (2012).
5. Xiao, Z.-D. *et al.* Energy stress-induced lncRNA FILNC1 represses c-Myc-mediated energy metabolism and inhibits renal tumor development. *Nat. Commun.* **8**, 783 (2017).
6. Khaitan, D. *et al.* The melanoma-upregulated long noncoding RNA SPRY4-IT1 modulates apoptosis and invasion. *Can. Res.* **71**, 3852–3862 (2011).
7. Zhang, J. *et al.* Screening and characterisation of sex differentiation-related long non-coding RNAs in Chinese soft-shell turtle (*Pelodiscus sinensis*). *Sci. Rep.* **8**, 8630 (2018).
8. Liu, M. *et al.* LncRNA-MEG3 promotes bovine myoblast differentiation by sponging miR-135. *J. Cell. Physiol.* jcp.28469 (2019). https://doi.org/10.1002/jcp.28469.
9. Wang, Y. *et al.* Overexpressing lncRNA LAIR increases grain yield and regulates neighbouring gene cluster expression in rice. *Nat. Commun.* **9**, 3516 (2018).
10. Faghihi, M. A. *et al.* Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of β-secretase. *Nat. Med.* **14**, 723–730 (2008).
11. Chiu, H.-S. *et al.* Pan-cancer analysis of lncRNA regulation supports their targeting of cancer genes in each tumor context. *Cell Rep.* **23**, 297-312.e12 (2018).
12. Lee, J. T. Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to the epigenome. *Genes Dev.* **23**, 1831–1842 (2009).
13. Mondal, T. *et al.* MEG3 long noncoding RNA regulates the TGF-β pathway genes through formation of RNA–DNA triplex structures. *Nat. Commun.* **6**, 7743 (2015).
14. Yoon, J.-H. *et al.* LincRNA-p21 suppresses target mRNA translation. *Mol. Cell* **47**, 648–655 (2012).
15. Kellis, M. *et al.* Defining functional DNA elements in the human genome. *Proc. Natl. Acad. Sci.* **111**, 6131–6138 (2014).
16. Ensembl. https://www.ensembl.org/info/data/ftp/index.html.
17. Jiang, S. *et al.* An expanded landscape of human long noncoding RNA. *Nucl. Acids Res.* **47**, 7842–7856 (2019).
18. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789 (2012).
19. Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigó, R. & Johnson, R. Towards a complete map of the human long non-coding RNA transcriptome. *Nat. Rev. Genet.* https://doi.org/10.1038/s41576-018-0017-y (2018).
20. The FAANG Consortium *et al.* Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol.* **16**, 57 (2015).
21. Giuffra, E., Tuggle, C. K. & FAANG Consortium. Functional Annotation of Animal Genomes (FAANG): Current Achievements and Roadmap. *Annu. Rev. Anim. Biosci.* **7**, 65–88 (2019).
22. Foissac, S. *et al.* Multi-species annotation of transcriptome and chromatin structure in domesticated animals. *BMC Biol.* **17**, 108 (2019).
23. Muret, K. *et al.* Long noncoding RNA repertoire in chicken liver and adipose tissue. *Genet. Sel. Evol.* **49**, 6 (2017).
24. Abril, J. F. Comparison of splice sites in mammals and chicken. *Genome Res.* **15**, 111–119 (2005).
25. Brown, W. R. A., Hubbard, S. J., Tickle, C. & Wilson, S. A. The chicken as a model for large-scale analysis of vertebrate gene function. *Nat. Rev. Genet.* **4**, 87–98 (2003).
26. Food and Agriculture Organization of the United Nations. http://www.fao.org/home/en/.
27. NCBI Resource Coordinators. Database Resources of the National Center for Biotechnology Information. *Nucl. Acids Res.* **45**, D12–D17 (2017).
28. Fang, S. *et al.* NONCODEV5: a comprehensive annotation database for long non-coding RNAs. *Nucl. Acids Res.* **46**, D308–D314 (2018).
29. Li, A. *et al.* ALDB: a domestic-animal long noncoding RNA database. *PLoS ONE* **10**, e0124003 (2015).
30. FR-AgENCODE · functional annotation of livestock genomes. http://www.fragencode.org/.
31. Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25**, 1915–1927 (2011).
32. Luo, S. *et al.* Divergent lncRNAs regulate gene expression and lineage differentiation in pluripotent cells. *Cell Stem Cell* **18**, 637–652 (2016).
33. Kambara, H. *et al.* Regulation of interferon-stimulated gene BST2 by a lncRNA transcribed from a shared bidirectional promoter. *Front. Immunol.* **5**, 676 (2015).
34. Modarresi, F. *et al.* Inhibition of natural antisense transcripts in vivo results in gene-specific transcriptional upregulation. *Nat. Biotechnol.* **30**, 453–459 (2012).
35. Beltran, M. *et al.* A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition. *Genes Dev.* **22**, 756–769 (2008).
36. Gibbons, H. R. *et al.* Divergent lncRNA GATA3-AS1 Regulates GATA3 Transcription in T-Helper 2 Cells. *Front. Immunol.* **9**, 2512 (2018).
37. Dhir, A., Dhir, S., Proudfoot, N. J. & Jopling, C. L. Microprocessor mediates transcriptional termination of long noncoding RNA transcripts hosting microRNAs. *Nat. Struct. Mol. Biol.* **22**, 319–327 (2015).
38. Odom, D. T. Control of pancreas and liver gene expression by HNF transcription factors. *Science* **303**, 1378–1381 (2004).
39. Lizio, M. *et al.* Systematic analysis of transcription start sites in avian development. *PLoS Biol.* **15**, e2002887 (2017).
40. Le Béguec, C. *et al.* Characterisation and functional predictions of canine long non-coding RNAs. *Sci Rep* **8**, 13444 (2018).
41. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE project. *Genome Res.* **22**, 1760–1774 (2012).
42. Wucher, V. *et al.* FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Res* gkw1306 (2017). https://doi.org/10.1093/nar/gkw1306.
43. Consortium, Gte. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
44. McCarthy, F. M. *et al.* Chickspress: a resource for chicken gene expression. *Database* **2019**, baz058 (2019).
45. Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).
46. Kryuchkova-Mostacci, N. & Robinson-Rechavi, M. A benchmark of gene expression tissue-specificity metrics. *Brief. Bioinform.* **18**, 205–241 (2016).
47. Schechtman, E. & Yitzhaki, S. On the proper bounds of the Gini correlation. *Econ. Lett.* **63**, 133–138 (1999).
48. Huminiecki, L., Lloyd, A. T. & Wolfe, K. H. Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases. *BMC Genomics* **4**, 31 (2003).
49. Simmerman, H. K. B. & Jones, L. R. Phospholamban: protein structure, mechanism of action, and role in cardiac function. *Physiol. Rev.* **78**, 921–947 (1998).

50. Barber, R. D., Harmer, D. W., Coleman, R. A. & Clark, B. J. GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues. *Physiol. Genomics* **21**, 389–395 (2005).
51. Hezroni, H. *et al.* Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Reports* **11**, 1110–1122 (2015).
52. Kern, C. *et al.* Genome-wide identification of tissue-specific long non-coding RNA in three farm animal species. *BMC Genomics* **19**, 684 (2018).
53. Muret, K. *et al.* Long noncoding RNAs in lipid metabolism: literature review and conservation analysis across species. *BMC Genomics* **20**, 882 (2019).
54. Wei, W., Pelechano, V., Järvelin, A. I. & Steinmetz, L. M. Functional consequences of bidirectional promoters. *Trends Genet.* **27**, 267–276 (2011).
55. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).
56. Freeman, T. C. *et al.* A gene expression atlas of the domestic pig. *BMC Biol.* **10**, 90 (2012).
57. Corney, B. P. A. *et al.* Regulatory architecture of the neuronal cacng2/Tarpγ2 gene promoter: multiple repressive domains, a polymorphic regulatory short tandem repeat, and bidirectional organization with Co-regulated lncRNAs. *J. Mol. Neurosci.* **67**, 282–294 (2019).
58. Prokopec, S. D. *et al. Identifying TCDD-resistance genes via murine and rat comparative genomics and transcriptomics* (2019). https://doi.org/10.1101/602698.
59. Bester, A. C. *et al.* An Integrated Genome-wide CRISPRa Approach to Functionalize lncRNAs in Drug Resistance. *Cell* **173**, 649-664.e20 (2018).
60. Guil, S. & Esteller, M. Cis-acting noncoding RNAs: friends and foes. *Nat. Struct. Mol. Biol.* **19**, 1068–1075 (2012).
61. Nakaya, H. I. *et al.* Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription. *Genome Biol.* **8**, R43 (2007).
62. Guil, S. *et al.* Intronic RNAs mediate EZH2 regulation of epigenetic targets. *Nat. Struct. Mol. Biol.* **19**, 664–670 (2012).
63. Khalil, A. M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *PNAS* **106**, 11667–11672 (2009).
64. McHugh, C. A. *et al.* The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature* **521**, 232–236 (2015).
65. Ha, M. & Kim, V. N. Regulation of microRNA biogenesis. *Nat. Rev. Mol. Cell Biol.* **15**, 509–524 (2014).
66. Rodriguez, A. *et al.* Requirement of bic/microRNA-155 for normal immune function. *Science* **316**, 608–611 (2007).
67. Maarouf, M. *et al.* Identification of lncRNA-155 encoded by MIR155HG as a novel regulator of innate immunity against influenza A virus infection. *Cell. Microbiol.* **21**, 1 (2019).
68. Cheng, L.-C., Pastrana, E., Tavazoie, M. & Doetsch, F. miR-124 regulates adult neurogenesis in the subventricular zone stem cell niche. *Nat. Neurosci.* **12**, 399–408 (2009).
69. Maiorano, N. & Mallamaci, A. Promotion of embryonic cortico-cerebral neuronogenesis by miR-124. *Neural Dev* **4**, 40 (2009).
70. Mele, M. *et al.* The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015).
71. Jenne, C. N. & Kubes, P. Immune surveillance by the liver. *Nat. Immunol.* **14**, 996–1006 (2013).
72. Cortes, M. *et al.* Metabolomics discloses donor liver biomarkers associated with early allograft dysfunction. *J. Hepatol.* **61**, 564–574 (2014).
73. Badawy, A.A.-B. Kynurenine pathway of tryptophan metabolism: regulatory and functional aspects. *Int. J. Tryptophan Res.* **10**, 117864691769193 (2017).
74. Lam, F. & Clifford-Mobley, O. Cholesterol Synthesis Defects. in *Disorders of Steroidogenesis* (eds. Rumsby, G. & Woodward, G. M.) 137–146 (Springer International Publishing, London, 2019).
75. Desert, C. *et al.* Multi-tissue transcriptomic study reveals the main role of liver in the chicken adaptive response to a switch in dietary energy source through the transcriptional regulation of lipogenesis. *BMC Genomics* **19**, 187 (2018).
76. Pevny, L. H., Sockanathan, S., Placzek, M. & Lovell-Badge, R. A role for SOX1 in neural determination. *Development* **125**, 1967–1978 (1998).
77. Zhao, Y. Control of hippocampal morphogenesis and neuronal differentiation by the LIM homeobox gene Lhx5. *Science* **284**, 1155–1158 (1999).
78. Morales, E. E. *et al.* Homeogene emx1 is required for nephron distal segment development in zebrafish. *Sci. Rep.* **8**, 18038 (2018).
79. Di-Poı, N. Distinct roles and regulations for hoxd genes in metanephric kidney development. *PLoS Genet.* **3**, 15 (2007).
80. Zakany, J., Darbellay, F., Mascrez, B., Necsulea, A. & Duboule, D. Control of growth and gut maturation by *HoxD* genes and the associated lncRNA *Haglr*. *Proc. Natl. Acad. Sci. USA* **114**, E9290–E9299 (2017).
81. Suhrie, K. *et al.* Neonatal lung disease associated with TBX4 mutations. *J. Pediatrics* **206**, 286-292.e1 (2019).
82. Zhang, W. *et al.* Spatial-temporal targeting of lung-specific mesenchyme by a Tbx4 enhancer. *BMC Biol.* **11**, 111 (2013).
83. Karolak, J. A. *et al.* Complex compound inheritance of lethal lung developmental disorders due to disruption of the TBX-FGF pathway. *Am. J. Hum. Genet.* **104**, 213–228 (2019).
84. Potter, S. S. Single-cell RNA sequencing for the study of development, physiology and disease. *Nat. Rev. Nephrol.* **14**, 479–492 (2018).
85. Plassais, J. *et al.* A point mutation in a lincRNA upstream of GDNF is associated to a canine insensitivity to pain: a spontaneous model for human sensory neuropathies. *PLoS Genet.* **12**, e1006482 (2016).
86. Jehl, F. *et al.* Chicken adaptive response to low energy diet: main role of the hypothalamic lipid metabolism revealed by a phenotypic and multi-tissue transcriptomic approach. *BMC Genomics* **20**, 1033 (2019).
87. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
88. Dobin, A. GitHub—alexdobin/STAR. https://github.com/alexdobin/STAR (2019).
89. Scott, E. Y. *et al.* Identification of long non-coding RNA in the horse transcriptome. *BMC Genomics* **18**, 511 (2017).
90. de Goede, O. M. et al. Long non-coding RNA gene regulation and trait associations across human tissues. Preprint at https://doi.org/10.1101/793091v1 (2019).
91. GitHub—tderrien/FEELnc: FlExible Extraction of LncRNA. https://github.com/tderrien/FEELnc.
92. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
93. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
94. The FANTOM Consortium *et al.* Interactive visualization and analysis of large-scale sequencing datasets using ZENBU. *Nat Biotechnol* **32**, 217–219 (2014).
95. NONCODE. http://www.noncode.org/index.php.
96. UCSC Genome Browser Home. https://genome.ucsc.edu/index.html.
97. NCBI. ftp://ftp.ncbi.nlm.nih.gov/.
98. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **12**, 323 (2011).
99. Koster, J. & Rahmann, S. Snakemake—–a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).

100. Snakemake/1000RNASeq_chicken/calling · master · bios4biol/workflows. *GitLab* https://forgemia.inra.fr/bios4biol/workflows/tree/master/Snakemake/1000RNASeq_chicken/calling.
101. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).
102. Szklarczyk, D. *et al.* STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucl. Acids Res.* **43**, D447–D452 (2015).
103. Chou, C.-H. *et al.* miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucl. Acids Res.* **46**, D296–D302 (2018).
104. Wong, N. & Wang, X. miRDB: an online resource for microRNA target prediction and functional annotations. *Nucl. Acids Res.* **43**, D146–D152 (2015).
105. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare. *F1000Res* **9**, 304 (2020).
106. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).

Acknowledgements

The data were collected in the frame of projects that received financial support from the French National Agency of Research (FatInteger project ANR-11-SVS7; ChickStress project, ANR-13-ADAP) and from the European Union's H2020 program under Grand Agreement No. 633531 (Feed-a-Gene project). FJ and KM are Ph.D. fellows supported by the Brittany region (France) and the INRAE Animal Genetics division. The authors also thank the staff of the INRA experimental poultry unit (UE1295 PEAT, Nouzilly, France) for producing and rearing animals, and the technicians of the research units for helping to measure birds. Finally, the authors thank S. Leroux, F. Boissel and M. Bellier for helping with RNA extraction.

Author contributions

F.J., K.M., H.A., E.G., S.D., S.F. and S.L. conceived the idea. F.J., K.M., Mo.B., C.D., F.P., T.Z. and S.L. participated to the set-up of the experimental designs and sample collection; Mo.B. and C.D. carried out all RNA extractions; D.E. generated RNA-seq libraries and sequencing; C.K. performed bioinformatics pre-processing of the RNA-seq data of 5 T dataset. M.B. carried out bioinformatics processing of the data of 21 T and re-processing of the data of 5 T dataset. P.D. generated the annotation in GRCg6a coordinates. F.J. and K.M. carried out all analyses. F.J., K.M., T.D., T.Z. and S.L. conceived and/or participated in the statistical analyses; M.B., C.D. and L.L. realized and interpreted PCR and RT-PCR analysis. F.J., K.M., P.D., H.A., E.G., S.D., S.F., T.D., F.P,. T.Z,. C.K. and S.L. drafted the manuscript. F.J. and S.L. drew the images used in the figures. All authors helped to draft the manuscript and read and approved the final version.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-77586-x.

Correspondence and requests for materials should be addressed to C.K. or S.L.

Reprints and permissions information is available at www.nature.com/reprints.

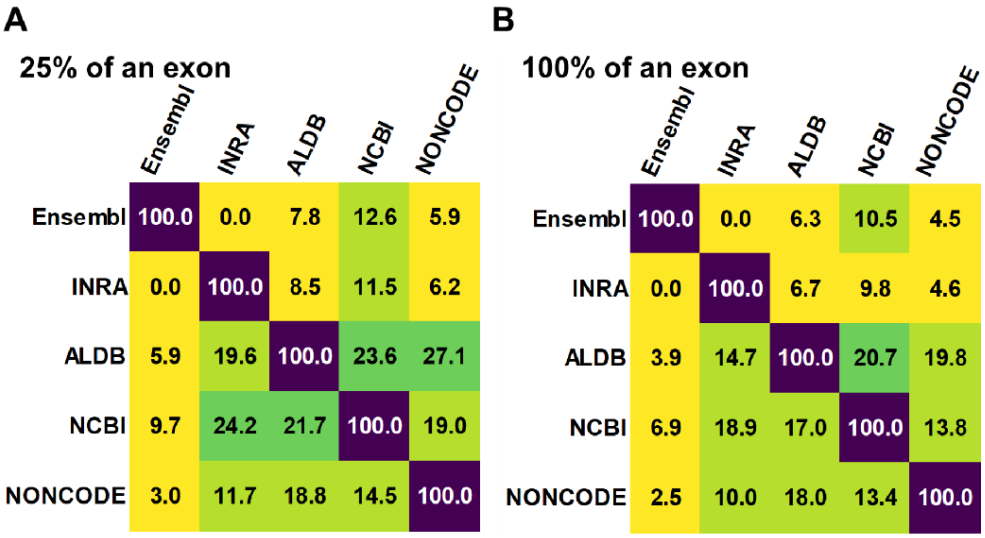
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

An integrative atlas of chicken long non-coding genes and their annotations across 25 tissues

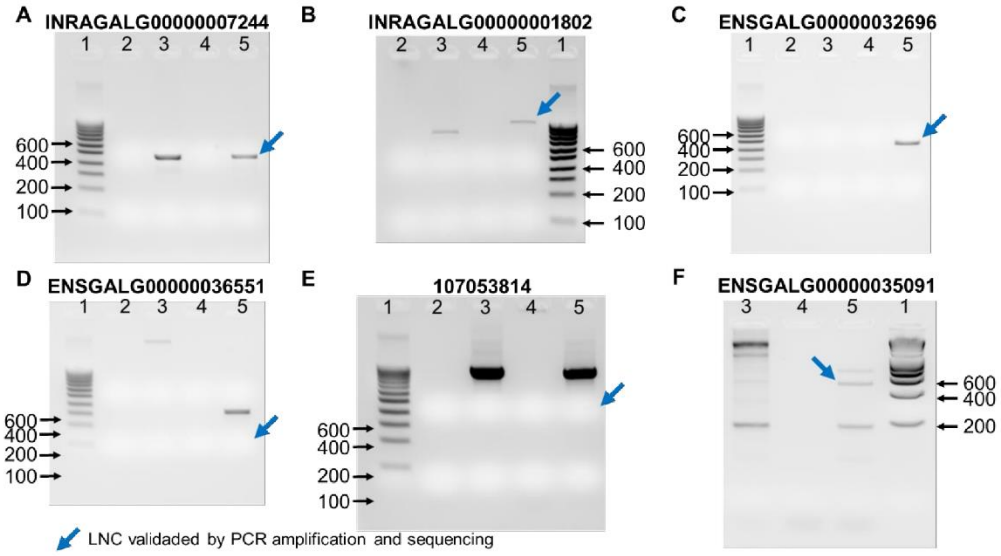
Frédéric Jehl[1,$], Kévin Muret[1,$], Maria Bernard[2,$], Morgane Boutin[1], Laetitia Lagoutte[1], Colette Désert[1], Patrice Dehais[2], Diane Esquerré[3], Hervé Acloque[4], Elisabetta Giuffra[5], Sarah Djebali[6], Sylvain Foissac[4], Thomas Derrien[7], Frédérique Pitel[4], Tatiana Zerjal[5], Christophe Klopp[2,*] and Sandrine Lagarrigue[1,*]

[1]PEGASE UMR 1348, INRA, AGROCAMPUS OUEST, 35590 Saint-Gilles, France
[2]SIGENAE Platform, INRA, 31326 Castanet-Tolosan, France
[3]GENOTOUL Platform, INRA, 31326 Castanet-Tolosan, France
[4]GenPhySE UMR 1388, INRA, INPT, ENVT, Université de Toulouse, 31326 Castanet-Tolosan, France
[5]GABI UMR 1313, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France
[6]IRSD, Université de Toulouse, INSERM, INRA, ENVT, UPS, Toulouse, France.
[7]IGDR UMR 6290, Univ Rennes, CNRS, 35000, Rennes, France

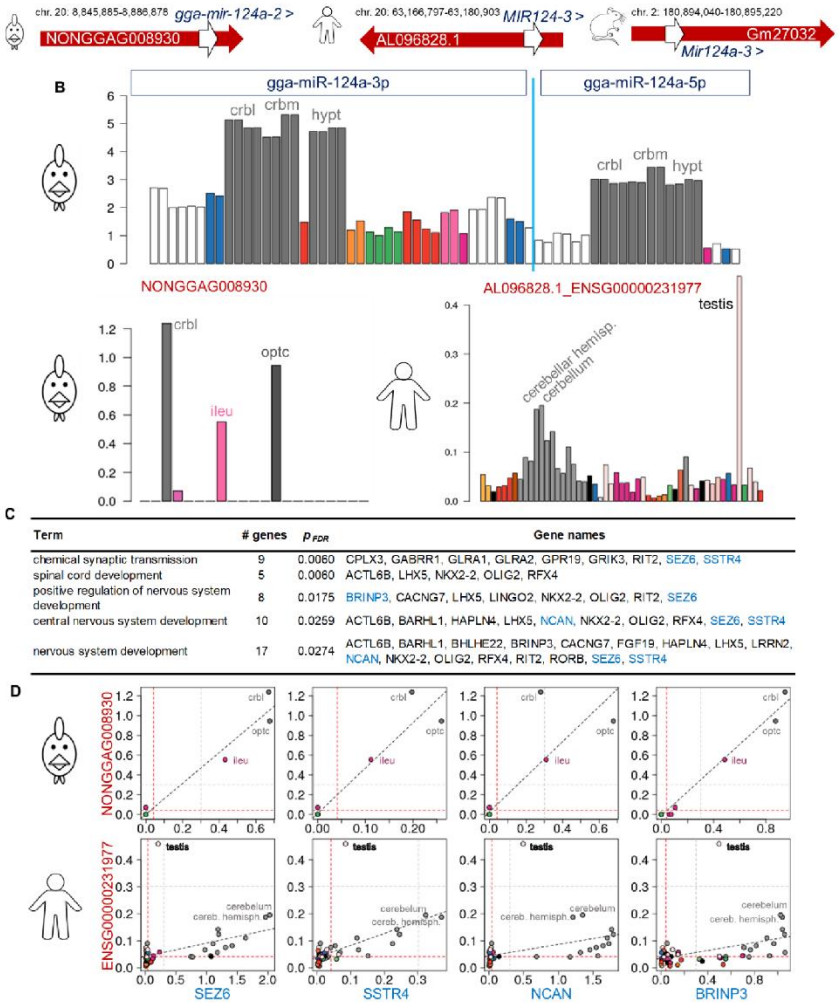
[*]Corresponding author



Supplementary Figure S1. Heatmap of the overlap between databases expressed in % of LNC (in line) shared among databases (in column), using 25% of an exon **(A)** and 100% of an exon **(B)** overlap



Supplementary Figure S2. Experimental validation of 6 LNC presented in main text. **(A)** INRAGALG00000007244, divergent of *PXDC1*. **(B)** INRAGALG00000001802, host of mir-155. **(C)** ENSGALG00000032696, divergent of *LHX5*. **(D)** ENSGALG00000036551, divergent of *TBX4*. **(E)** 107053814, divergent of *HOXD3*. **(F)** ENSGALG00000035091, divergent of *SOX1*. Number above each gel correspond to: 1: ladder; 2: PCR negative control; 3: PCR using DNA; 4: PCR using RNA; 5: PCR using cDNA. Arrows next to the ladder indicate the size of the ladder's fragments. Blue arrows next to the band indicate the LNC that was validated by PCR amplification and sequencing.



Supplementary Figure S3. Conservation of genomic location and function of a miR host LNC across species. (A) NONGGAG008930 hosts gga-mir-124a-2 and might be conserved in human (AL096828.1) and mouse (Gm27032). **(B)** gga-mir-124a-2 (left) and its host LNC, NONGGAG008930 (middle), are mostly expressed in immunity-related tissues in chicken, similarly to AL096828.1 in human (right). Gga-mir-124a-2 expression is expressed in $\log_{10}$(FPKM + 1), NONGGAG008930 and AL096828.1 expressions are expressed in $\log_{10}$(TPM + 1) **(C)** Top 5 enriched KEGG terms supported by more than 5 genes associated to the PCG correlated to NONGGAG008930. PCG in blue are used in next panel. **(D)** Co-expression of four PCG from previous panel with NONGGAG008930 in chicken (top) or AL096828.1 in human (bottom).

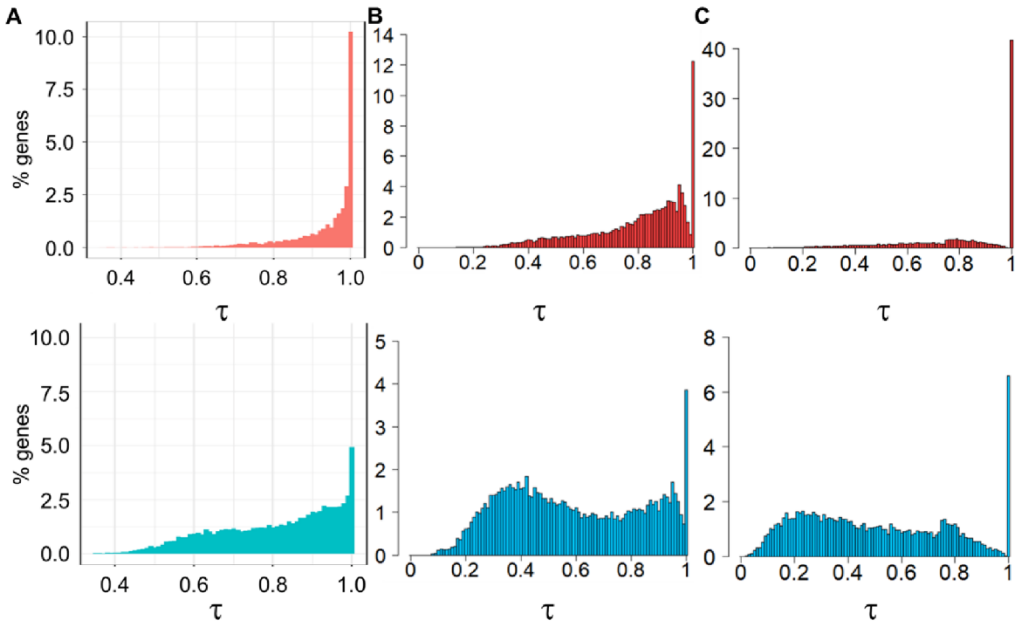
Among the 89 PCG, we found 3 targets of the chicken or human miRNAs:

hsa-miR124-3p: *RFX4* detected as a target (target score = 98),

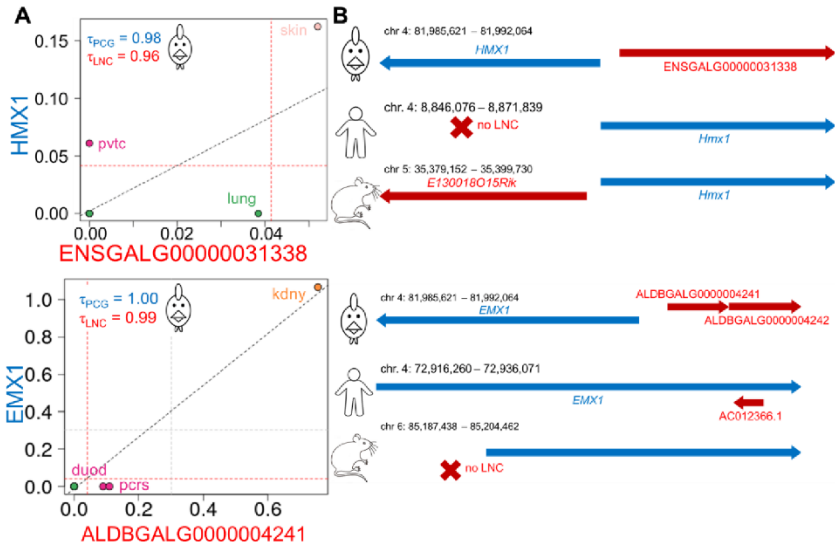
has-miR124-5p: no target detected (target score ≥ 80),

gga-miR-124a-3p: *RFX4* (target score = 93) was also detected as a target, along with *NOL4* (target score = 92).

gga-miR-124a-5p: *NOL4* (target score ≥ 80) was also detected as a target, along with *RIT2* (target score ≥ 80).



Supplementary Figure S4. Distribution of the τ values in dog (A), and in chicken 21T (B) and chicken 5T (C) projects. Top, in red, τ values associated to LNC. Bottom, in blue, τ values associated to PCG



Supplementary Figure S5. Tissue expression in chicken **(A)** for the 2 LNC:PCG pairs with similar genomic configurations between the two species **(B)** not presented in Main Text Figure 6. **(A)** log10(TPM+1) expression of the LNC (X-axis) and the PCG (Y-axis) (top) for chicken across the 21 tissues of the 21T dataset for the two chicken divergent pairs for which both members are tissue-specific in the same tissue. **(B)** Genomic configuration in three species of the LNC:PCG pairs in chicken, human and mouse. In red the LNC, in blue the PCG. **NB:** Corresponding LNC in human was either not found (for LNC:*HMX1* pair) or absent from the GTEx expression dataset (for LNC:*EMX1* pair).

f) précision sur l'obtention de la classification des LNC au niveau des gènes

Comme nous l'avons vu plus haut, la classification des LNC par rapport au PCG le plus proche se fait d'abord à l'échelle des transcrits (chaque transcrit du LNC est classifié par rapport à chaque transcrit du PCG). Il est donc tout à fait possible pour un LNC donné d'avoir des transcrits classifiés de manière différente à des PCG différents. Pour passer à la classification à l'échelle des gènes, il peut être nécessaire de décider d'un ordre de préséance que nous détaillons ici.

Nous avons d'abord favorisé les configurations géniques, pour lesquelles la possibilité d'une erreur de modélisation est plus évidente, aux configurations intergéniques. En effet, on peut penser qu'un LNC dont le modèle chevaucherait un intron d'un PCG sur le même brin risque d'être en fait une partie de ce PCG. Rappelons que nous avons éliminé les LNC dont ne serait-ce qu'un exon chevauchait sur le même brin un exon d'un PCG. Ainsi, parmi les géniques, nous avons favorisé les *sens* aux *antisens* à cause du risque manifeste que le LNC soit un exon mal modélisé du PCG. Les *antisens* ont pour leur part un intérêt biologique, puisque différents travaux ont mis en évidence leur capacité à réguler le PCG qu'ils chevauchent sur l'autre brin[281,284–286,324]. Ensuite, parmi les intergéniques, nous avons favorisé les *divergents* aux *même-brins*. En effet, bien que les « mêmes-brins » puissent être en réalité un bout d'un PCG mal rabouté (voir aussi Figure 4 de l'introduction de la thèse), la configuration divergente autorise des hypothèses de co-régulation dans le cas d'une co-expression[283,288] qui sont intéressantes que le gène soit en effet un LNC ou bien un PCG mal modélisé. De plus, la configuration même-brin ne doit pas être automatiquement associée à des cas suspects : on trouve aisément dans les génomes bien annotés (chez l'humain par exemple) des PCG fiables et contiguës en *même-brins*. En tout état de cause, nous nous sommes proposés de signaler clairement dans notre annotation étendue les LNC classifiés comme « même-brin » avec un PCG. Enfin, vient la configuration *convergente* qui ne semble pas présenter de risque particulier relatif à l'annotation, et pour laquelle il n'y a pas dans la littérature d'hypothèses biologiques particulières autour d'une éventuelle co-expression.

2. Le RNA-seq pour la détection de SNP fiables : potentiel pour la détection de variants affectant les régions codantes et l'étude de l'expression allèle-spécifique (article 2)

a) contexte et objectifs

Le travail précédent a fait notamment usage des LNC modélisés dans le cadre d'une précédente thèse[177] à partir d'une grande quantité de données RNA-seq. Nous avons ici souhaité utiliser cette masse de données RNA-seq à une autre fin : la détection de SNP dans les régions exprimées, et par suite, la détection des génotypes associés à ces SNP chez les individus. Ces informations sont nécessaires à l'étude de l'expression allèle-spécifique dans le cadre de la recherche de gènes candidats causaux de l'efficience alimentaire dans la partie III.

Le RNA-seq est actuellement la méthode de choix pour l'étude des transcriptomes. Cette méthode a diverses applications, que nous avons détaillées dans l'introduction de la présente thèse. Grâce à son étape de séquençage, le RNA-seq permet d'accéder à la séquence des transcrits, et donc aux polymorphismes qu'ils portent. La vaste majorité de ces polymorphismes portés par les transcrits sont issus de polymorphismes portés également par l'ADN. L'*editing* pourrait être un obstacle à ce projet mais comme indiqué dans l'introduction, l'application de filtres classiques quant aux « *mismatches* » (différences entre séquence du *read* et génome de référence) pour l'alignement des *reads* de RNA-seq font que les *reads* porteurs d'événements d'*editing* sont très largement éliminés. Ainsi les études appliquant les procédures classiques de traitement de données de RNA-seq ne reportent que quelques dizaines à centaines d'événements d'*editing* dans les tissus analysés quelles que soient les espèces étudiées : poule[186,187], souris[180,188] et humain[180,188]. Il est important de noter qu'avec le RNA-seq, et contrairement au DNA-seq, le nombre de *reads* supportant la détection d'un SNP peut fortement varier d'une position à l'autre, à cause du niveau d'expression des gènes. Si à l'échelle de la population, les SNP peuvent être détecté avec une bonne fiabilité en cumulant l'information portée par les *reads*, les génotypes par individu doivent pour leur part être considérés avec précaution. Dans un premier temps, nous avons cherché des critères permettant de sélectionner les SNP ayant des génotypes nous paraissant suffisamment fiables.

Les données de RNA-seq ne sont que peu, voire pas, utilisées pour la détection de SNP (certainement pour des raisons techniques que nous aborderons plus loin) alors qu'elles représentent un volume de données conséquent et en accès libre. Ces données se sont en effet

accumulées ces dernières années, dans le cadre de projets divers, essentiellement centrés sur l'analyse de différentiels d'expressions dans des populations variées. Malgré cela, l'ordre de grandeur du nombre de SNP détectés par RNA-seq à partir d'un tissu et d'une population donnée n'est pas bien connu vu le faible nombre d'études sur le sujet. Ainsi l'étude présentée ici avait trois objectifs. (*i*) D'abord, fournir des ordres de grandeur du nombre de SNP et de génotypes détectables, en utilisant des données de RNA-seq issues de 10 populations de poules (lignées de pontes et de chairs, commerciales et expérimentales). (*ii*) Ensuite, donner une vue générale des conséquences prédites de variants localisés dans les régions codantes, et notamment du nombre de variants prédits comme délétères (au sens « perte de fonction » définie par le consortium gnomAD[345]), en travaillant avec un grand nombre de populations avec plusieurs dizaines d'individus par population, ce que ne permet en général pas de faire le séquençage DNA-seq, encore trop coûteux. Ces variations « perte de fonction » sont largement étudiées pour identifier celles qui sont responsables d'un phénotype par un impact sur la séquence de la protéine ou pour mieux comprendre les fonctions des gènes affectés. Cependant, elles sont contre-sélectionnés dans les populations naturelles et donc rares, mais peuvent être détectés avec un certain nombre d'échantillons dans une population donnée. Chez les espèces modèles comme l'humain, les régions codantes du génome ont beaucoup été séquencées par séquençage de l'exome (WES). Ainsi, le consortium gnomAD[345] a analysé 125 748 exomes humains provenant de sources publiques, et a identifié 443 769 variations « perte de fonction » prédites avec un haut degré de confiance. (*iii*) Enfin, donner des ordres de grandeur du nombre de gènes pouvant être analysés par ASE, ainsi que du nombre de SNP analysables par gène (c'est à dire hétérozygotes). Nous avons également illustré la possibilité d'utiliser les SNP détectés par RNA-seq pour explorer la diversité génétique de ces 10 populations en utilisant des jeux de SNP prédits comme ayant des conséquences plus ou moins délétères. Ces travaux ont été conduits dans le cadre de la présente thèse et avec un étudiant de master 2 que j'ai co-encadré. Ils ont également fait l'objet d'une collaboration avec l'équipe INRAE Sigenae.

b) matériels et démarche

Pour évaluer le potentiel du RNA-seq dans la détection de SNP, nous avons commencé par comparer les SNP détectés par RNA-seq à ceux détectés par DNA-seq dans deux populations indépendantes de 15 et 8 poules pour lesquelles les deux types de données étaient disponibles strictement sur les mêmes échantillons biologiques (même tissus et même individus). Pour cette comparaison entre DNA-seq et RNA-seq, nous nous sommes concentrés sur les régions exprimées du génome, afin d'assurer une comparaison équitable des deux techniques, et avons

appliqué les filtres proposés par GATK (à l'origine de la suite d'outils utilisée pour la détection des variants) pour assurer la qualité des SNP détectés par RNA-seq. Nous avons commenté ces filtres proposés au vu des résultats obtenus, puisque ces filtres ne sont pas encore standardisés en RNA-seq. Ensuite, nous nous sommes intéressés aux génotypes associés aux SNP détectés par RNA-seq en étudiant leur concordance entre DNA-seq et RNA-seq, afin de s'assurer de leur qualité. En effet, par opposition au DNA-seq, rappelons que la difficulté majeure inhérente aux données de RNA-seq est l'absence d'homogénéité des profondeurs des *reads* d'un locus à l'autre, puisque cette profondeur dépend de l'expression des transcrits. Suite à ces analyses, nous avons proposé deux filtres, que nous verrons plus loin, permettant une concordance entre DNA-seq et RNA-seq suffisamment élevée. Forts de cette procédure, nous avons exploré des données de RNA-seq déjà disponibles sur 10 populations dans lesquelles les SNP ont été détectés et filtrés avec les critères que nous avons proposés précédemment. L'identification des variants ayant un fort effet sur une protéine a été faite à l'aide de l'outil VEP (*Variant Effect Predictor*)[346] d'Ensembl et l'évaluation du nombre de gènes analysables par ASE a été faite à l'aide de scripts idoines. L'exploration de la diversité génétique des 10 populations a été faite à l'aide des outils du package R *SNPRelate*[347] en utilisant des jeux de SNP dont les conséquences avaient été prédites.

c) résultats

D'abord, nous avons observé qu'à régions équivalentes (c'est-à-dire dans les exons exprimés des gènes exprimés), 91% des SNP détectés en RNA-seq le sont également en DNA-seq, et ce en utilisant des données de 2 populations différentes (avec n = 15 et n = 8). La majeure partie des SNP détectés par RNA-seq sont par ailleurs situés dans des introns, dans lesquels le RNA-seq garde un bon taux de vrais-positifs. Concernant les 9% de SNP détectés par RNA-seq et pas par DNA-seq, environ 50% sont situés dans des « *SnpCluster* », c'est-à-dire des groupes de 3 SNP ou plus dans une fenêtre de 35 nucléotides. C'est aussi le cas pour 43% des SNP détectés uniquement par DNA-seq, et pour 20% des SNP détectés par les deux méthodes. Le reste des SNP détectés uniquement par RNA-seq peuvent être expliqués par d'autres facteurs de moindre importance : proximité avec une jonction exon-exon (5.1% *versus* 3.66% pour les SNP détectés par les deux méthodes), régions de faibles complexité (3.7% *versus* 1.9%), présence des régions non-traduites, aux extrémités des transcrits (3 à 4% *versus* 3 à 10%). Nous avons ensuite cherché à sélectionner des SNP pour lesquels les génotypes dans la population étudiée pouvaient être considérés comme fiables. Nous avons montré que la concordance entre les génotypes détectés par RNA-seq (sans utilisation d'aucun filtre) et ceux détectés par DNA-seq

était relativement élevée (environ 90%), et qu'imposer des critères de pourcentage de génotypes détectés (ici, 50%) et de génotypes soutenus par un certain nombre de *reads* (ici, au moins 20% des génotypes soutenus par au moins 5 *reads*) permettait d'assurer une concordance d'environ 95% tout en conservant environ 40% des SNP détectés. D'une population à l'autre, nous avons ainsi détecté entre 1 300 000 et 4 100 000 de SNP dans le seul tissu qui leur était commun, le foie, et 9.9 millions en utilisant l'ensemble des données à notre disposition. De même, nous avons détecté selon les populations entre 300 000 et 1 000 000 ayant des génotypes répondant à nos critères dans la population avec les données de foie, avec en moyenne 560 000 par population. L'étude des conséquences prédites des variants a montré, comme attendu, que la plupart d'entre eux (97%) étaient situés dans des régions non-codantes, et que parmi ceux qui étaient situé dans des régions codantes, 62% étaient des variants de type synonymes. Au final, seules 0.54% de toutes les conséquences prédites avaient un effet délétère sur la protéine codée, ce qui représente tout de même 101 755 variants, compte tenu du nombre de populations (10) et d'animaux analysés (337 individus). L'évaluation du nombre de gènes analysables par ASE nous a montré que 72% des PCG et 59% des LNC exprimés à TPM ≥ 1 étaient analysables, en utilisant uniquement les SNP détectés dans leurs exons, avec tout de même des variations entre populations (par exemple pour les PCG, entre 48 et 89% des gènes sont analysables selon la population). L'étude de la diversité génétique des populations à l'aide de jeux de SNP ayant des conséquences prédites plus ou moins délétères reste quant à elle à finaliser.

d) discussion et conclusion

La détection de SNP par RNA-seq ouvre, nous l'avons vu, la voie à l'analyse des nombreuses données RNA-seq accumulées au fil des ans sous un nouvel angle. Le RNA-seq permet de détecter avec un taux élevé de vrais positifs des variants dans les régions exprimées du génome. Il devient alors possible d'étudier à grande échelle les conséquences des variants présents sur les régions codantes exprimées, même celles qui ont un impact parmi les plus sévères et sont donc rares. Moyennant ensuite quelques filtres, que nous proposons, il est possible de détecter des génotypes fiables grâces à cette méthode, permettant disposer de SNP hétérozygotes qui pourront être utilisés pour des analyses d'expression allèle-spécifique. Ces travaux effectués sur différentes populations ayant des origines variées fournissent pour la première fois des ordres de grandeurs concernant le nombre de SNP détectables à l'aide de données de RNA-seq.

e) article, en cours de relecture par les co-auteurs

Cet article est en cours de relecture par ses co-auteurs : **Frédéric Jehl***, Fabien Degalez*, Maria Bernard*, Frédéric Lecerf, Manon Coulée, Colette Désert, Sophie Leroux, Laetitia Lagoutte, Diane Esquerré, Benham Abasht, David Gourichon, Thierry Burlot, Michelle Tixier-Boichard, Bertrand Bed'hom, Tatiana Zerjal, Frédérique Pitel, Christophe Klopp and Sandrine Lagarrigue. RNA-seq for reliable variant detection: interest for coding variant characterization and allele-specific expression. Soumission prévue à *Frontiers in Genetics* début 2021.
*co-premiers auteurs

Il a fait l'objet d'un « e-poster » (présenté à distance) : **F. Jehl**, M. Bernard, F. Degalez, F. Lecerf, M. Coulee, T. Zerjal, F. Pitel, C. Klopp and S. Lagarrigue. Genomic SNP detection by RNA-seq: lessons from multi-tissue & multi-population data analysis in chickens. 20è Journées Ouvertes de Biologie, Informatique et Mathématique (JOBIM),
Montpellier (France), du 30 juin au 2 juillet 2020.

Il est reproduit ci-après.

RNA-seq for reliable variant detection: interest for coding variant characterization and allele-specific expression

Frédéric Jehl*, Fabien Degalez*, Maria Bernard*, Frédéric Lecerf, Manon Coulée, Colette Désert, Sophie Leroux, Laetitia Lagoutte, Diane Esquerré, Benham Abasht, David Gourichon, Thierry Burlot, Michelle Tixier-Boichard, Bertrand Bed'hom, Frédérique Pitel, Tatiana Zerjal, Christophe Klopp** and Sandrine Lagarrigue**

*co-firsts authors

**corresponding authors

Abstract

Background: In addition to their classic usages, RNA-seq data that accumulated in the last 10 years provides a yet-unexploited resource of SNP in numerous individuals from different populations. SNP detection by RNA-seq is particularly interesting for livestock species since whole genome sequencing is expensive and tools for whole exome sequencing is unavailable. These SNPs and their genotypic frequencies could be used to characterize variants affecting protein function, in particular the rare ones, and, combined with expression, to study *cis*-regulation by allele-specific expression (ASE) analysis. However, these data have highly variable expression levels and GATK filters for SNP detection are not yet standardized, making challenging SNP and genotype detection by RNA-seq.

Results: Using GATK filters suggested for RNA-seq and two independent chicken population for which RNA- and DNA-seq data were available on the exact same samples (n = 15 and n = 8), we show in expressed regions that 91% of the SNP detected by RNA-seq are also detected in DNA-seq 20X, and characterized the remaining 9% SNP. We then studied the concordance of genotypes (GT) computed with DNA-seq versus RNA-seq according to two factors (GT call-rate and number of reads supporting the GT) and proposed thresholds for these factors that lead to a 95% of concordance. Using these thresholds on 744 RNA-seq samples collected on 10 chicken populations, we found 9.9M of SNP in total, with on average 560,000 SNP per population with a GT call rate ≥ 50% and 340,000 with a MAF ≥10%. We then show in chicken species that such RNA-seq data *i*) can detect numerous variant effects on protein with a severe predicted impacts despite their rarity in each population (more than 100,000 SNP), and *ii*) allow to analyse on a large scale, *cis*-regulation by using the ASE approach (with on

average 72% of protein-coding and 60% long non-coding genes with expression ≥ 1TPM being analysable). We also illustrated the possibility to perform population genetic analysis with such SNPs detected exclusively in expressed regions with more or less functionally deleterious impact, alternatively to the analysis based on the standard LD or HD SNP microarrays.

Conclusions: This work shows that RNA-seq data can be used with good confidence to detect SNP and associated GT within across various populations and used them for different types of analysis.

Background

RNA-seq is currently the method of choice to study transcriptome expression in replacement of gene chips [1] in various research topics such as genic regulations [2], transcriptome response to environmental changes [3], diseases [4], lines with contrasted phenotypes [5]. However, the RNA-seq method has other, more specific applications that take advantage of its sequencing step such as transcript and gene modelling [6][article 1de la thèse], allele-specific expression analysis, which combines SNP at RNA levels and expression in the investigated gene to assign an expression level for each chromosome [7], or the analysis of RNA editing, a phenomenon resulting in a SNP observed at the RNA level, while the position is homozygous at the DNA level [8]. Another application of RNA-seq based on its sequencing step, is to access to the SNP of the expressed part of the genome, as proposed by Piskol *et al*. in 2013 [9]. It is particularly interesting in non-model species like livestock species in which no exome tools have been developed as alternative of DNA-seq data which remains costly to generate and store. In this context, RNA-seq presents different advantages. First, the number of RNA-seq data already available is incomparably higher than the number of DNA-seq data, whichever the species (chicken, pig, cow and other non-model species) since these data accumulated for the last several years and keep accumulating for transcriptomic studies in different populations of the species. Moreover, in each population and study, RNA-seq data are usually available on several dozen individuals since these studies focus on two and often more conditions, with a dozen (in general unrelated) individuals per condition (often 8 to 12) allowing to better catch, in a given population, variants with low frequency compared to those detected by DNA-seq which is used on a few individuals per population because of its cost. Second, RNA-seq data allow studying the variations in the coding regions of the expressed genome, in which it is easier to predict the functional impact of SNPs. Interestingly, some of these SNPs can have a strong effect on protein function, leading to a “loss of function”. These loss-of-function variants are extensively studied

to identify those that are responsible for a phenotype through an impact on the protein sequence [10]. In addition, they represent a powerful source of information to better understand the functions of the affected genes [10]. However, such "loss of function" SNPs are counter-selected in natural populations and therefore rare but can be detectable with a certain number of samples in a given population. In well-known model-species like human, these coding regions are accessible using exome sequencing (WES), as shown by the recent work of the Genome Aggregation Database (gnomAD, published in 2020) [10]. This consortium analysed 125,748 human exomes (and much less whole genomes: 15,708) from public sources, and identified 443,769 high-confidence predicted loss-of-function variants, defined in the work of gnomAD as being either stop gained (non-sense variants), frameshift or splice site variants. For non-model species such as livestock species, in which WES method is usually not available, RNA-seq can thus fulfil the same objective, with similar advantage of WES that produces a lower volume of data, thus facilitating storage and decreasing its cost [11]. Third, in addition to the SNPs they could detect, RNA-seq data provide expression values of the loci from which these SNPs are located, allowing to study allele-specific expression as we previously exposed, and hence, *cis*-regulation at large scale, in multiple tissues and multiple populations. Fourth, the transcribed regions are overall well spread in the genome and much more numerous than previously thought since the ENCODE project highlighted thousands of long noncoding genes [12]. RNA-seq data should therefore provide sets of numerous and well spread SNPs in the genome. Finally, these data could be used to analyse the genetic diversity of populations with a different point of view compared to the SNP chips, by offering various sets of SNP with functional impacts more or less severe.

Nevertheless, RNA-seq is currently not often used for SNP detection in coding regions despite the advantages just mentioned. Indeed, SNP and genotype detection by RNA-seq presents three main challenges. First, the transcriptome is composed of mature transcripts (i.e. spliced), making more difficult the mapping of RNA-seq reads overlapping exon-exon junctions, compared to DNA-seq reads [13]. However, mapping methods for RNA-seq data seems to be well mastered since the first paper of Piskol *et al.* in 2013 [9], even though it is important to remain cautious for the SNP detected close to the exon-exon junctions [7, 14]. Second, RNA editing, by definition, could have been a strong limitation for SNP detection by RNA-seq since it introduces SNP at RNA level, which are absent at the DNA level. Nevertheless, as we discuss later, RNA editing has such features that this phenomenon only slightly impedes the detection of reliable SNPs by RNAseq in standard conditions. Third, transcripts have highly variable

expression levels, leading the read depth of a few reads to a few thousands reads, contrarily to the DNA-seq that offers a homogeneous read depth along the genome. Coding and noncoding transcripts can be expressed at vastly different levels, from few copies to millions of copies per cell, in different cell types and developmental or physiological stages. Moreover, the transcriptome is also composed for a small portion of immature transcripts under processing (composed of exons and introns). These introns are more variable in sequence but less supported by reads compared to the exons [11]. This variation in depth from a gene to another, and within a gene, between intron and exon, constitutes a major challenge for the detection of SNPs, and, more importantly, for the detection of genotypes at the individual level by RNA-seq. Indeed, reliable SNP detection at the population level benefits of the accumulation the information born by the reads across individuals (see Figure 1).

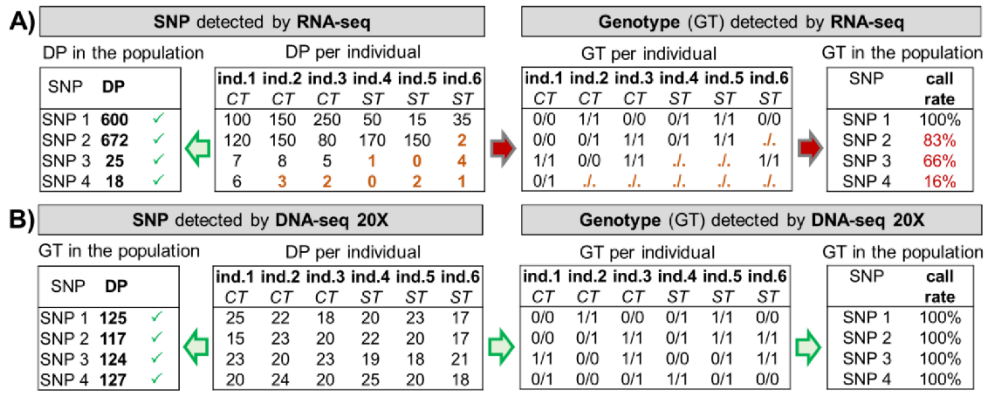


Figure 1: Toy example of illustrating the need for read depth (DP) filters in RNA-seq and the difference with DNA-seq. **A)** *"DP per individual"*: DP data per sample showing RNA-seq read depth for 4 SNPs in 6 samples from 2 experimental settings (Control CT and Stress ST). SNP#1 shows good read coverage across all samples, as well as SNP#2 except in ind. #6. SNP#3 is borne by a gene that has a lower expression in ST compared to CT, hence a lower DP for the samples in this condition. SNP#4 was borne by an overall lowly expressed gene. *"DP in the population"*: at the population scale, the existence of the four SNPs is known, albeit with different depths. *"GT per individual"*: Detection of GT per sample using the DP data per sample. The GT are reliable for all individuals for SNP#1, as well as SNP#2, except for ind. #6 that had low coverage, resulting in unknown genotype "./." (brown GT and DP). For SNP#3, GT are detected only in the individuals of the CT group. For SNP #4, only one GT is detected. *"GT in the population"*: at the population scale, the call-rate for the genotypes (CR) shows that only SNP#1 has a GT in all samples, and that SNP#4 cannot be used to compute meaningful statistics. **B)** DNA-seq offers a homogeneous coverage of the genome (20X in our case), all the SNP are therefore detected at the individual level with a DP of 20 reads on average, and at the population level with a DP of 6×20=120

reads on average. All the genotypes can therefore be computed at the individual level, resulting in a call rate of 100% for every SNP.

This last point might explain why only few studies have used RNA-seq data for variant detection since the first publication in 2013. A consequence of that fact is that neither the magnitude of the number of SNPs that could be detected in the frame of a "classical" RNA-seq study, nor of the percentage of individuals with a genotype, a prerequisite before computing allelic frequencies, are known. To our best knowledge, since Piskol *et al.* in 2013 [9], less than a dozen works proposed large-scale SNP detection tools from RNA-seq data ([15–21]) and no studies have applied these tools to answer to a biological question. The reference tools evolving very rapidly, these studies have successively tested different tools, and among them, only Adetunji *et al.*, 2019 [21] used the most recent tools proposed by ENCODE for RNA-seq data, i.e., STAR [22] for the mapping and GATK [23] for variant detection. Three of the above-mentioned studies were interested in the concordance of the SNP and the genotype detection between RNA-seq and DNA-seq, the latter being the gold standard for SNP detection. However, these studies used only a few samples (from 1 to 4) and had not at their disposal both the RNA-seq and DNA-seq data on the same tissues of the same individuals.

In this context, this work aims at detecting SNP from RNA-seq data in the chicken. The first goal was to put up a procedure allowing the detection of SNPs and genotypes (GT) from RNA-seq data, using the reference tools (STAR for read mapping and GATK for SNP detection). We tested the SNP reliability according to the three filters suggested by the GATK team and comparing the detected SNPs with those obtained using DNA-seq data. This comparison work was realized in two independent chicken populations for which RNA-seq and DNA-seq data were available on the very same biological samples (i.e. the same tissue of the same individuals): liver samples from a layer population (Population A, $n = 15$) and a broiler population (Population B, $n=8$). We also studied the effects of adding other tissues in the number of SNP detected.

Because a large proportion of SNP detected by RNA-seq was reliable, we further applied this procedure to ten different chicken populations with three objectives. These populations were: a wild Red Jungle Fowl population, Fayoumi line, six commercial and experimental laying hen lines and two commercial and experimental broiler lines. The three goals were to (*i*) provide an estimation of the number of detectable SNP and GT using RNA-seq data in each population,

(*ii*) present an overview of the predicted consequences of the SNP in the coding regions, in particular the number of high-confidence predicted loss-of-function variants, as defined in the work of gnomAD and finally (*iii*) give an overview of the RNA-seq potential for allele-specific expression (ASE) analysis, by estimating the number of genes that could be analysed for ASE and the number of SNP per gene. Finally we illustrated the possibility of using RNA-seq data to explore genetic diversity between populations using these 10 populations and two lists of RNA-seq SNP set with variable percentage of SNP with a severe predicted protein consequence.

Results and discussion

SNPs detected by RNA-seq are concordant with those detected by DNA-seq, they are mostly located in introns, and their number depends on the number of tissues used.

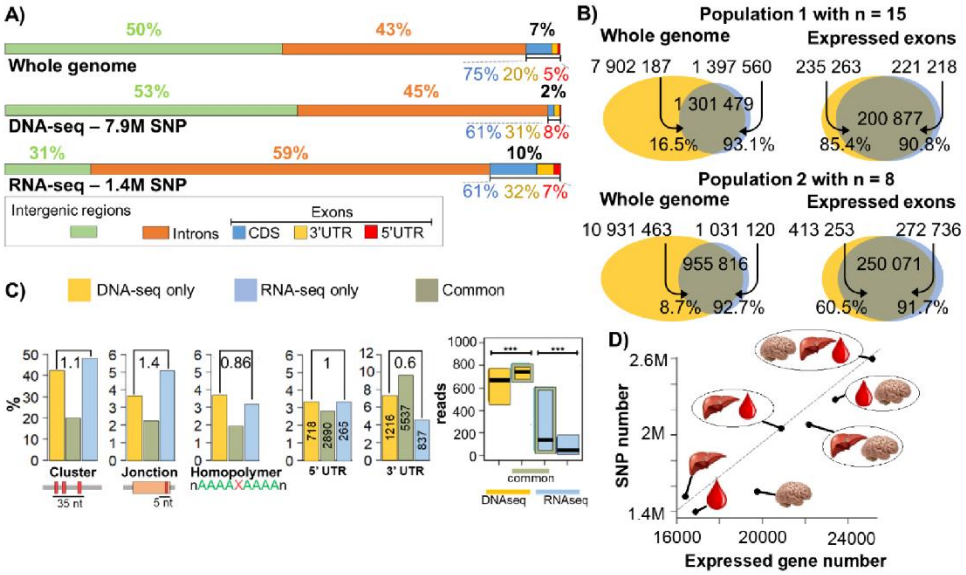


Figure 2: RNA-seq and DNA-seq differences and common features. A) Number of SNPs detected by DNA-seq (yellow set), RNA-seq (blue set) and by both methods (grey set) at the whole genome level (left) and the level of the expressed exons (right) in two independent population A (n=15) and B (n=8). **B)** Distribution of reads supporting SNP at the population level, detected by DNA-seq only, RNA-seq only and both methods (common) in DNA-seq and RNA-seq data, and percentages of these SNPs in junctions, homopolymers, SNP clusters (i.e. 3 or more SNPs in a sliding window of 35 bp, as per GATK definition), 5' and 3' UTR in population A. **C)** Distribution of the length of the genomic features across the genome (top) and of the SNPs detected by DNA-seq (middle) and RNA-seq (bottom) across these genomic features. **D)** Evolution of the number of detected SNPs as a function of the number of

expressed genes using one tissue alone or groups of tissues. Tissues used were liver (figured as a cartoon liver), blood (figured as a blood drop) and hypothalamus (figured as a brain).

We compared the repartition of the SNPs detected by DNA-seq and RNA-seq to the proportions of the genome covered by different features (Figure 2A). The chicken genome is composed at equal parts of intergenic and genic sequences, with 43% of introns and 7% of exons. These exons are composed at 75% of coding sequence (CDS), at 20% of 3'-UTR and a 5% of 5'-UTR. As expected, DNA-seq SNP were well distributed in the genome (53% in intergenic regions, 45% in introns and 2% in exons) but were located in lower proportion in exonic regions, that are under selection pressure, than the proportion of these regions in the genome (2% vs 7%). With RNA-seq, we expected that most of the SNP would be in exonic regions, which represent expressed regions. However, the majority of the detected SNPs were located in intronic regions (59%) and in intergenic region (31%).The detection of SNPs in intronic regions can be explained by the presence of unspliced transcripts, certainly lowly expressed compared to spliced transcripts, but sufficiently to be supported by reads, and by the low selection pressure on these regions compared to the exons (54 SNP per intron versus 6.5 SNP per exon). The SNPs located in "intergenic regions" are likely to be located in new genes or part of genes (particularly 3'- and 5'-UTRs) yet-unmodelized. Within exons, the proportion of SNP in 3'-, 5'-UTR and CDS were similar (32%, 7%, 61%) between RNA-seq and DNA-seq, but different from the proportion of these regions in the genome (20%, 5%, 75%) showing clearly a lower selection pressure in 3'UTR regions than in CDS regions.

We detected SNPs using either RNA-seq or DNA-seq data obtained from the liver of the same 15 laying hens. We found 7,902,187 biallelic SNP using the DNA-seq data filtered with the standard criteria of GATK (see Methods). We therefore consider these SNP as reliable. Using the RNA-seq data filtered with some of the filters suggested by GATK (see Methods and comments below), we found 1,397,560 SNP (Figure 2B). As expected, the number of SNPs detected with RNA-seq is much lower than in DNA-seq, since only variants present in transcribed regions are detected.

To provide a meaningful comparison of the two methods, we used the SNPs detected in the expressed exons of the expressed genes, assessed using RNA-seq (see Methods). Indeed, even though the current RNA sequencing technologies allow us to determine whether or not a gene is expressed, it is difficult to know which of its transcript(s) do undergo transcription. Using the metric described in Material & Methods, we detected 162,145 expressed exons among the

147,474 exons of the 16,814 expressed genes (on average 8.8 exons per gene). In these exons, more than 85% of the 235 263 SNP detected with DNA-seq were also detected by RNA-seq (Figure 2B), meaning that RNA-seq allows to detect almost all the SNPs located in expressed exons, and has therefore a good sensibility at this level. In the second analysed population B, which was composed of only 8 samples, we found that the number of SNP detected by both methods corresponded only to 60% of the SNP detected by DNA-seq. This difference in sensibility for the RNA-seq is likely due to the number of samples (8 in population B versus 15 in population A), that causes an overall number of reads bearing information for each position to low for the SNP to be detected and to satisfy the quality criteria at the population level (see Figure 1).

In addition, among the 221,128 SNP detected by RNA-seq in population A, 91 % were also detected by DNA-seq 20X showing that most of the SNP detected by RNA-seq are reliable, and that RNA-seq also has a good precision at this level. Interestingly, we obtained the same percentage of precision (91%) using the second population B. These results are consistent with the work from Guo *et al*. (2017) [24], who compared the percentage of SNP detected using RNA-seq versus exome sequencing and found around 85% of concordance. Regarding the 9% of RNAseq-specific SNP, we analysed different factors that could underlie their detection, in order to reduce or at least signal some of them as potential false-positives to be treated with caution. With this aim in view, we compared the impact of these different factors between the RNAseq-specific SNP *versus* the DNA-seq specific SNP and the SNPs detected by the both methods (Figure 2C). We consider the latter two sets as reliable since DNA-seq are now routinely used for SNP detection with the well-proven GATK filters. **First**, we observed that a large proportion of RNA-seq specific SNP (48.35%) were considered as belonging to a "SNP cluster" (i.e. 3 or more SNPs in a sliding window of 35 bp, as per GATK definition). We found a similar percentage for DNA-seq specific SNP (43%) (Figure 2C). This filter is one of the three filters proposed by GATK for RNA-seq SNP detection, but with the remark that these filters are not definitive and must be validated by users. It is noteworthy that 20% (i.e. 1/5th) of the SNPs detected by both methods are also located in "SNP clusters". Therefore, considering that, first, GATK's filters for RNA-seq are declared by their authors as needing refinement, second, that this filter is not used for DNA-seq data, and third, that a large share of SNPs detected by DNA-seq (only or in common with RNA-seq) fall into this category, we decided not to remove those SNPs from our dataset, but only to flag them as belonging to a so-called SNP cluster. The remaining RNA-seq specific SNP can be explained by other factors of lowest impact: (*i*) 5.07% were located at 5 bp or less of an exon-exon junction, *versus* 3.66% and 2.22% for those

detected only by DNA-seq and those in common, respectively. The ratio "RNA-seq specific / DNA-seq specific" significantly greater than 1 (1.4, $p \leq 10^{-20}$, χ^2 test) was expected since RNA-seq deals with spliced transcripts (Figure 2C) and their mapping by the aligner is more complicated and more error-prone. Since most of them are also observed in DNA-seq, we consider that the SNP in the vicinity (i.e. 5 bp) of the junctions can be kept, but that they must be targeted and validated by another technique. (*ii*) 3.2% were located in low complexity regions, defined as repetition of at least 5 identical nucleotides, versus 3.7% for the ones detected only by DNA-seq and 1.93% for the ones detected by both methods (Figure 2C). (*iii*) 3.4% and 4.3% were observed in 5'UTR and 3'UTR regions respectively with a lowest presence of 3'-UTR SNPs compared to the ones detected only by DNA-seq (0.6, $p \leq 10^{-20}$, χ^2 test) (Figure 2C). This may be due to the fact that mature transcripts undergo the action of exonucleases that degrade their 3' extremities, making them unsequensable [25]. (*iv*) To finish, another factor responsible for these RNA-seq specific SNP is RNA editing, although it is unlikely, according to the literature, that most of the remaining SNPs are due to this mechanism. In mammals, in which editing is well studied, Adenosine-to-inosine (A-to-I) editing due to ADAR1 and ADAR2 enzymes is the most common form of editing and mostly occur in inverted pairs of Alu repeats [26], which the chicken genome contains a homologous family called CR1 [27]. These editing events tend to occur in clusters, a phenomenon called hyper-editing that introduces ≥ 20 mismatches in the sequencing reads [28], that are therefore discarded by the aligner either because of a multi-mapping or no mapping. The prevalence of editing is still discussed: RNA editing seems relatively rarely detected when working with classical mapping filters as shown in human [29, 30], in mouse [31] and in chicken with less than 100 events [8, 32, 33], and frequently detected when working in repeated regions and rescuing unaligned reads [34]. Finally, we observed that the SNPs detected only by one method were covered by significantly less reads (either of RNA- or DNA-seq) than the SNPs detected by both methods (Figure 2C).

Using samples from two blood and hypothalamus collected on the same 15 animals, we studied the effect of detecting the SNPs in more than one tissue. Results are displayed in Figure 2D. We detected 1,397,560 SNPs in the liver (as previously stated), 1,511,690 in the blood and 1,557,416 in the hypothalamus, while 16,814 genes were expressed in the liver, 16, 346 in the blood, and 19,733 in the hypothalamus. When using combinations of two or three tissues, the number of detected SNP increased in relation with the number of expressed genes (spearman correlation = 0.96, $p = 6 \times 10^{-4}$). Using several tissues, one can generate a file (VCF file)

comprising all the SNPs detected in the population, allowing us to cumulate the information of all tissues and ensure their reliability by associating each of them with the tissue in which they were covered by the highest number of RNA-seq reads (see Material and Methods and Supplementary Figure 1 [*en cours de réalisation, voir également Figure 7 de la présente thèse*]).

Importance of genotype call rate (CR) and read depth at the individual scale for selecting SNP with enough reliable genotypes for calculating genotype & allele frequencies

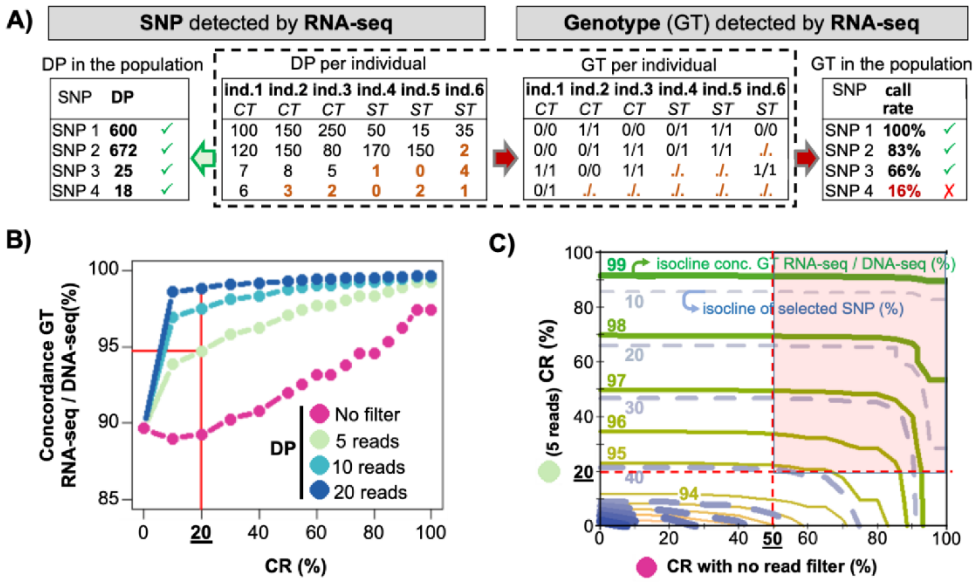


Figure 3: The passage from SNP to GT necessitates a read depth threshold. A) Toy example of illustrating the need for read depth (DP) filters with the same legend as in Figure 1 *Right part*: the call-rate for the genotypes (CR) shows that SNP#4 cannot be selected to compute allelic frequencies, the CR threshold of 50% not being met. **B)** Evolution of the percentage of concordance between RNA-seq and DNA-seq as a function of the percentage of GT in the population CR supported by at least 0 reads (purple curve), or 5 (light green), 10 (light blue) and 20 (dark blue). **C)** Contour plot showing the values of percentage of concordance between RNA-seq and DNA-seq (solid green lines) and of the percentage of selected SNP compared to the original set (dashed blue lines) depending on the values of CR with no read filter (x-axis) and the CR with at least 5 reads ("(5 reads) CR", y-axis). Red surface: SNP selected after filtering on (5_{reads}DP)CR ≥ 20% and a CR ≥ 50%.

While reliable SNPs can be detected in the population thanks to some individuals that bear them, it doesn't necessarily mean that there are enough reads in each individual to allow the

genotyping tool to assign a genotype ("GT") to each individual. This was exemplified in Figure 1 in the introduction by the brown cases (SNPs #3 and #4): there are enough reads in each individual to detect the existence of the SNP at the population level, but there are not enough reads in some individuals to estimate their genotype. In Figure 1, this was the case for the individuals of the "stress" group (ind. #4, #5 and #6) for the SNP#3 or most of the individuals of the population (except ind. #2 to #5) for the SNP#4. These cases are quite frequent in practice because of the variability of gene expression from one individual to another, especially when different conditions impacting transcriptomes are analysed or also when a SNP is located in an intron of an immature transcript (by definition, weakly expressed compared to the associated mature transcript). Therefore, the call rate (CR), defined as the % of individuals with a genotype (GT) in the population, can be very highly variable (e.g. from 33% to 100% in Figure 3A) from a SNP to another, depending on the number of reads observed in each individual (henceforth called "read depth", DP). With 20X DNA-seq data, most of the SNP have a CR close to 100%, as schematized in Figure 1.

The reliability of GT at the individual level depends also on the DP supporting the GT. This reliability was tested through the concordance between the RNA-seq GT and the DNA-seq GT. Here, we aimed at studying the relationship between 4 parameters: (*i*) the number of reads supporting the GT (the GT depth, called DP), (*ii*) the GT concordance between RNA-seq and DNA-seq, which is the criteria for the reliability of the GT, (*iii*) the CR, and finally, (*iv*) the number of SNP kept depending on the values of these three parameters. First, we studied conjointly the effects of the criteria "CR" and "DP supporting the GT" on the concordance between RNA-seq and DNA-seq (Figure 3B) and found a concordance relatively high (around 90%) without threshold for the DP (purple line). This concordance increases to around 95% for a CR $\geq$ 20% of genotypes with a DP $\geq$ 5 reads (light green line) and over 97% for a CR $\geq$ 20% of genotypes with a DP $\geq$ 10 reads (light blue line). We then studied in Figure 3C the impact of the CR of genotypes without a DP threshold (x-axis) conjointly of the CR of genotypes supported by $\geq$ 5 reads, (noted (5_{reads}DP)CR in y-axis), on the concordance between RNA-seq and DNA-seq (Figure 3C, solid green isoclines). We also provide the number of SNP conserved according to the different criteria (Figure 3C, dashed blue isoclines). Interestingly, for a CR (x-axis) between 0% and 70%, only the (5_{reads}DP)CR has an effect on the concordance and the percentage of conserved SNP, hence the horizontal isoclines. Hence, we propose for our subsequent analysis on different RNA-seq datasets, two filters: (5_{reads}DP)CR $\geq$ 20% ensuring a concordance of almost 95% and around 40% of conserved SNP and a CR $\geq$ 50% ensuring a sufficient number of GT per SNP to calculate the allelic frequencies. We can note in this

population and most of the populations analysed in the next section that more than 98% of SNP with (5_{reads}DP)CR ≥ 20% have a CR ≥ 50% (Additional Data S1).

Number of detectable SNP and GT using RNA-seq data in 10 populations

Pop.	# ind.	# smpl.	# tiss.	Total SNP			Selected GT			Selected GT & MAF ≥ 10%			
				Liver[a]	Multi-tissues[b]	b / a	Liver[c]	Multi-tissues[d]	d / c	Liver[e]	Multi-tissues[f]	e / a	f / b
RJFh	18	72	3	–	2 646 463	–	277 194	583 914	2.11	151 204	324 447	–	0.12
Cobb	48	96	2	4 122 798	5 867 458	1.42	952 757	1 686 055	1.77	557 528	951 059	0.14	0.16
FLLL	32	64	2	1 833 854	3 416 944	1.86	537 688	1 114 302	2.07	368 109	714 065	0.20	0.21
A3A3	32	32	1	1 490 773	1 490 773	1	449 768	449 768	1	264 698	264 698	0.18	0.18
N4A3	64	104	2	1 308 551	2 172 690	1.66	391 955	740 806	1.89	243 790	449 249	0.19	0.21
RpRm	88	286	5	1 883 978	4 130 952	2.19	557 563	1 288 405	2.31	306 928	631 358	0.16	0.15
RMx6	19	19	1	–	2 223 852	–	–	718 531	–	–	483 236	–	0.22
FrAg	4	7	2	1 267 812	1 764 922	1.39	789 382	986 728	1.25	525 301	427 566	0.41	0.24
Lsnu	16	32	2	1 526 260	2 355 898	1.54	593 152	840 429	1.42	384 682	534 812	0.25	0.23
Fayo	16	32	2	1 349 391	2 089 712	1.55	498 272	701 654	1.42	288 432	396 381	0.21	0.19
Mean	–	–	–	1 847 927	2 815 966	1.66	560 859	911 059	1.78	343 408	517 687	0.22	0.20
Union	337	744	–	5 682 906	**9 949 072**	–	1 678 971	3 423 310	–	1 244 012	2 112 626	–	–
Intersection		–	–	294 528	288 484	–	**69 896**	85 127	–	**2 717**	2 017	–	–

Figure 4: Overview of the populations and number of SNPs per population retained at each step of the selection. Total SNP: SNP detected at the level population; Selected GT: SNP with at least 50% of genotypes and 20% of GT with reads ≥ 5 reads (see Figure 3); Selected GT and minor allele frequency (MAF) ≥ 10%. Union: union of SNP detected in at least one population. Intersection: Polymorphic SNP (i.e. with at least one ALT allele) in each of the 10 populations. Superscripts are used to show which ratio are presented.

As shown in Figure 4 that gives an overview of the SNP diversity of 10 chicken populations, we detected between 1.2M and 4.1M SNP per population using liver RNA-seq datasets. Using all the tissues available (1 to 5 tissues depending on the population), we detected between 1.5M and 5.9M SNP (see Figure 2D): consistently with our previous result, we observed a fold increase of ×1.39 to ×2.19 depending on the number of analysed tissues. Across the populations and using all the tissues, we found a grand total of 9.9M SNP having at least one alternative allele in at least one population (SNP union), and 288 484 SNP that had at least one alternative allele in each of the 10 populations (SNP intersection). We compared the SNP union and the SNP intersection sets with the SNPs listed in the reference dbSNP database [38] and those available on the 600K genotyping chip [39], respectively. The union of our SNP contains 25% (2 429 880) yet-unrepertoried SNPs in the dbSNP database (21M SNP) and the intersection of our SNP contains 2.4% (13 674 SNP) of the SNP of the 600K sequencing chip.

We then selected SNPs with genotypes (Figure 4, “Selected GT”) and found between around 450 000 and 1.7M SNP using all tissues, with a union of 3.4M SNP, much less than the 9.9M SNP observed previously. These results on 10 populations show that a large number of SNP

(two thirds) are detected at the population scale thanks to the accumulation of the reads across all the individuals of the population, but that within each individual, the read number is not sufficient to reliably determine a genotype. Nevertheless, the total number of SNP with GT remains high, in the order of magnitude of the million (3.4M) with an intersection of 85 127 SNP. In the liver, for which data were available in all but one population (RMx6), the union and intersection are of the same order of magnitude: 1.7M and 69 896 SNP, respectively. After selecting for a MAF (minor allele frequency) ≥ 10%, which corresponds to 4 observations of the allele in the smaller populations of 16 individuals and up to 18 observations for the largest one (88 individuals), the number of SNP is halved in all the populations, and as expected the intersection drastically goes down to 2 717 SNP, since this set corresponds to the SNP with a MAF ≥ 10% in each of the 10 populations.

Rare deleterious variants detection in the populations

We predicted the consequences of the 9 949 072 SNP detected in at least one population using the VEP tool. We found 18 679 930 consequences (VEP predicts the consequences of the SNP in each of the transcripts it affects), caused by 9 706 800 unique variants, affecting 24 279 genes. As expected, the vast majority of the SNP (97%, *i.e* 18 147 051) affected non-coding regions versus 2.5% (474 364) affecting coding regions, 0.3% (54 945) a splice site in a coding-gene and the rest a splice site in a non-coding gene (n = 3570, 0.02% of total). Among the SNPs affecting a coding-region, as expected, a majority (62%) were synonymous whereas 37% were missense variants. The rest affect start (n = 592, 0.1%) and stop codons (n = 2856, 0.6%), with in particular 2369 stop gained predictions (81% of all predictions affecting stops) affecting 1469 genes (n = 1806 variants). Among all these predictions, a focus was put on the predicted consequences with the most severe impact used by the gnomAD consortium [10]: start and stop loss, variant in the splice regions or stop gain. Noteworthy, the severity of the latter might depend on its position in the CDS: a mostly translated protein might be functional, while a heavily truncated one probably won't. Some tools like SIFT ("Sorting Intolerant from Tolerant") [35] also associate a severity score for missense variant, depending on the conservation across species of the affected amino-acid [36]. Since SNP detection using RNA-seq allows working on numerous individuals and populations, numerous variants affecting coding regions can be detected (474 364 in our case), and therefore even rare variants with "strong effect" predicted consequences can be observed. In our case, we listed 2 856 predictions affecting stops (319 stop lost, 167 stop retained and the 2369 stop gained evoked previously), 592 predictions affecting starts (all start lost), 54 945 predictions affecting a coding gene splice

site, and 40 228 SIFT-predicted deleterious missenses (i.e., 23% of all missense), for a grand total of 101 755 severe predicted impacts (0.54% of all consequences), corresponding to 66 743 SNP and 13 845 genes.

Focusing on SNP with genotypes (41 454 out of the 101 755), we found 23 952 variants that were present at both the heterozygous and homozygous states in at least one population which suggest they have mild effect, and 17 502 that were only present at the heterozygous state, which could mean that the homozygous state is strongly counter-selected, suggesting an important role for the gene.

Potential for allele-specific expression analysis in the populations

A)

expr.	biotype	SNP†	gene	S / g§	GT†	genes	S / g§	%₡	MAF†	genes	S / g§	%₡	Het ≥ 25% †	genes	S / g§	%₡
SNP in exons																
1 TPM	PCG	**185 075**	**11 696**	**16**	**148 567**	**10 559**	**14**	**90**	**92 060**	**9 275**	**10**	**79**	**70 066**	**8 371**	**8**	**72**
	LNC	76 006	3 005	25	49 743	2 435	20	81	29 774	2 094	14	70	19 210	1 774	11	59
0.1 TPM	PCG	218 562	14 112	15	158 422	11 703	14	83	97 590	10 177	10	72	72 809	8 935	8	63
	LNC	187 911	9 223	20	76 563	4 753	16	52	45 459	3 950	12	43	27 763	3 056	9	33
SNP in genes (i.e. exons + introns)																
1 TPM	PCG	1 183 140	11 696	101	466 929	10 797	43	93	260 271	9 784	27	84	167 787	8 227	20	71
	LNC	117 823	3 005	39	63 324	2 472	26	82	37 996	2 145	18	71	24 251	1 829	13	60
0.1 TPM	PCG	1 272 001	14 112	90	432 422	11 979	36	85	268 529	10 729	25	76	171 919	9 342	18	66
	LNC	272 934	9 223	29	98 925	4 848	20	53	59 169	4 069	15	44	35 996	3 173	11	34

B) *Exons only*, PCG – 1 TPM: details by population

Pop	SNP†	gene	S / g§	GT†	genes	S / g§	%₡	MAF†	genes	S / g§	%₡	Het ≥ 25% †	genes	S / g§	%₡
RJFh	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Cobb	305 494	11 796	26	236 251	11 301	21	96	137 925	10 846	13	92	105 186	10 484	10	89
FLLL	233 029	15 301	15	179 854	11 715	15	77	123 389	11 226	11	73	76 896	9 905	8	65
A3A3	169 349	11 065	15	132 290	10 335	13	93	78 562	9 003	9	81	61 311	8 199	7	74
N4A3	168 762	11 074	15	130 471	10 260	13	93	81 924	9 204	9	83	65 079	8 398	8	76
RpRm	150 546	10 580	14	119 325	9 751	12	92	66 519	7 486	9	71	35 028	5 115	7	48
FrAg	*155 302*	*12 782*	*12*	*141 264*	*11 599*	12	91	*95 861*	*9 804*	10	*77*	*88 413*	*9 555*	9	*75*
Lsnu	161 208	10 452	15	135 723	9 844	14	94	86 944	8 808	10	84	72 827	8 198	9	78
Fayo	136 907	10 518	13	113 357	9 669	12	92	65 354	7 825	8	74	55 787	7 113	8	68
Mean	**185 075**	**11 696**	**16**	**148 567**	**10 559**	**14**	**90**	**92 060**	**9 275**	**10**	**79**	**70 066**	**8 371**	**8**	**72**
CV (%)	**30**	**14**	**27**	**27**	**8**	**22**	**7**	**28**	**14**	**15**	**9**	**30**	**20**	**13**	**17**

Figure 5: Overview of the analysable genes for allele-specific expression in the liver for two biotypes (PCG and LNC) and at two gene expression levels (0.1 TPM and 1 TPM). **A)** Average across the 10 populations of the number of SNP and expressed genes containing at least one SNP satisfying different criteria. GT: SNP with genotypes satisfying CR filters. MAF: SNP from GT column with minor allele frequency (MAF) ≥ 10%. Het ≥ 25%: SNP from the MAF column with a frequency of the heterozygous ≥ 25%. Analysis of SNP located in the exons (top) or in genes (introns and exons) (bottom). **B)** Details by population for the PCG expressed at 1 TPM with at least one exonic SNP. † indicates a number of SNP. § Number of SNP per gene. ₡ Percentage of genes at each step versus initial number of expressed genes. CV: coefficient variation, i.e. standard deviation divided by the mean.

Allele-specific expression (ASE) analysis mingles both variant detection and expression analysis to quantify the expression level of both alleles of a gene, in general at a heterozygous SNP position, to test an eventual imbalance in the expression between the two chromosomes.

It is therefore necessary to observe at least one heterozygous variant in the expressed feature. Usually, the expression is evaluated using RNA-seq and the variants are detected using DNA-seq, which is expensive when working on a dozen or more individuals. Since we have shown that RNA-seq allows observing a large set of reliable SNP in expressed regions, we studied in this section, the potential for ASE analysis using only RNA-seq data. To this end, we counted for the ten chicken populations the number of genes having at least one SNP satisfying our criteria, with a MAF ≥ 10% and heterozygous in at least 25% of the population, allowing testing the ASE across individuals (Figure 5A). We also indicated the average SNP number per gene (column "S / g") to give an idea of the potential to test ASE along the gene, using all the SNP they bear. Results are presented in Figure 5A for two types of genes: the protein-coding genes (PCG) which are the best known genes, and the long non-coding RNA (LNC), a class of genes producing transcripts that do not code for a protein, which are increasingly considered as important gene expression regulators but that are also known to be less expressed than PCG [6, 12, 37]. This is the reason why we studied each class of genes with two different expression threshold: 0.1 TPM and 1 TPM. The first is classically used when working on LNC, since these genes globally have a lower expression level than PCG, that are usually studied with an expression threshold of 1 TPM. Finally, results in Figure 5A are presented either for SNP detected in the exons of the expressed genes (top) or for SNP detected in the whole gene (i.e. in the exons or the introns, bottom). It is better to work on exonic SNP, which belong to a spiced transcript, than on intronic SNP that have more chance to belong to an unspliced transcript, and for which the number of associated reads may not be representative due to the low "expression" of the introns. Interestingly, we show that the number of genes with at least one SNP are similar in both cases (exons only versus exons + introns), despite a lower number of SNP per gene when the SNP are only selected in exons. When working with exonic SNP (Figure 5A, bottom), there are on average 15 to 25 SNP per gene showing the possibility to test ASE along genes. This number is higher for LNC (20-25) compared to PCG (15-16) probably due to the lower selection pressure of LNC compared to PCG. Moreover, we show that 72% (8371) of PCG and 59% (1774) of LNC expressed at TPM ≥ 1 are analysable by ASE, i.e. approximately 10 000 genes that bear at least one SNP for which at least 25% of the individuals of the population are heterozygous, with on average 10 SNP per gene. Figure 5B shows the variability between the ten populations depending on the intra population heterozygosity frequency. For example, we observed that 53% (5600) for "RpRm" to 91% (10700) for "Cobb" of the PCG (TPM ≥ 1 TPM) are analysable for ASE, showing again that the latter line is overall more heterozygous than the former. The same tendencies regarding the percentage of genes that can be analysed were

observed for the PCG (TPM ≥ 0.1) and for LNC (both for TPM ≥ 0.1 and ≥ 1) (Additional File S2).

Use for diversity exploration in ten chicken populations

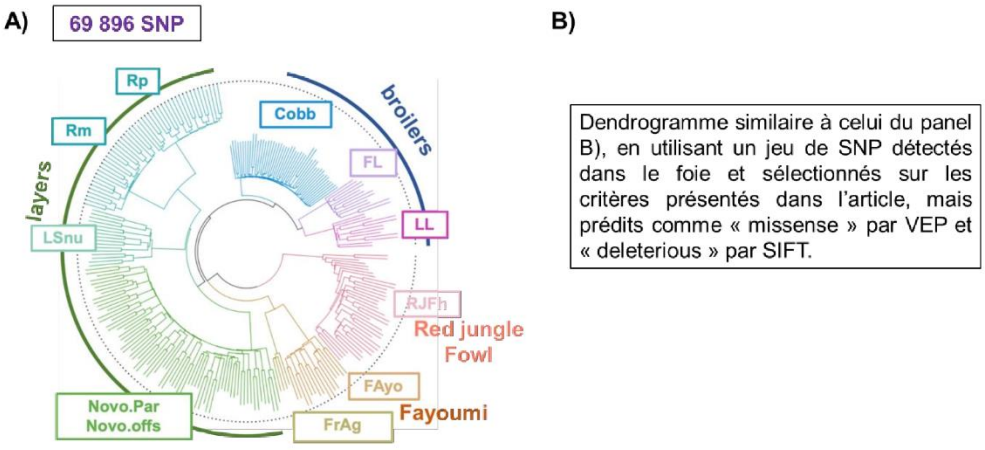


Figure 6: Overview of the diversity of the 10 populations using two sets of SNP with different predicted consequences. A) CAH using the 69 896 SNP intersection set (with GT, see Figure 4). **B)** CAH using the *X* [*analyse à réaliser*] SNP intersection set predicted as "missense" and "deleterious".

We studied the genetic links between the population using the genotypic frequencies of the intersection of the SNP with GT in the liver, i.e. the 69 876 SNP, which represent a set of SNP of a different nature than those used on genotyping SNP chips. Indeed, the latter are considered as having a neutral effect, while most the SNP present in our data are located in translated regions and affect to some extent the protein (from almost neutral synonymous to deleterious stop gained). To test whether using potentially deleterious SNP could change the classification, we used only the SNP predicted as "missense" by VEP and "deleterious" by SIFT (Figure 6B). The classifications with the 69 876 SNP (Figure 6A) is consistent with the history of the populations, indicating that these SNP detected by RNA-seq and their associated GT allow distinguishing different populations. The classification separated the RJFh (a Red Jungle Fowl population, used here to represent the "ancestral" line), then the layers (light red circle arc) and the broilers (light blue circle arc), and within these populations, the commercial (Novo.Parents and Novo.Offsrpings for the layers, Cobb for the broilers) and experimental (RpRm for the layers and FLLL for the broilers) ones. Within the experimental population, there is a clear distinction between two subpopulations that are divergent for a trait (feed efficiency for the

RpRm and body fat for the FLLL). The classifications with the *X* [*analyse à réaliser*] SNP (Figure 6B) revealed [*analyse à réaliser*].

Material and methods

Animals and tissues

For the comparison of the SNP detected by RNA-seq versus DNA-seq, we used two populations for which both data type were available on the same liver samples from the same birds. The first population ("population A") was composed of 15 birds from an experimental layer line of animals diverging for feed efficiency after a 40-year diverging selection, the R+ and R- [40]. The second population ("population B") was composed of 8 birds from an experimental broiler line of animals diverging for body fat content, the FLLL [41]. For the rest of the work, we used RNA-seq data from 10 populations: a red jungle fowl population (called RJFh); two broiler lines, a commercial one, the Cobb 500 (Cobb Vantress, called Cobb) and the FLLL presented previously; six layer lines: two brown eggs commercials ones, the Novo.Parents and Novo.Offspring (NOVOGEN), two experimental ones, the R+ and R- (RpRm) presented previously, the Naked Neck (LSnu), composed of dwarf chicken without feather around the neck, the FrAg, composed of leghorn chicken, and finally, the Fayoumi (FAyo), an Egyptian line. See Additional file 3 for the detail of the number of birds, the tissues and the number of samples.

Tissues sampling, RNA and DNA isolation and sequencing

For the FLLL and the RpRm, tissues sampling and RNA isolation were described in Muret *et al.*, 2017 [6] and Jehl *et al.*, 2019 [3], respectively. Raw data of both DNA-seq and / or RNA-seq are available on the ENA and SRA archives under accession numbers: PRJEB28745 (RpRm, Novo.Parents and Novo.Offsprings, RNA-seq); SRP079637 (FLLL, RNA-seq); PRJEB26695 (red jungle fowl, RNA-seq); PRJEB34341 (naked neck, RNA-seq); PRJEB34310 (Fayoumi, RNA-seq), ERP023985 (FrAg, RNA-seq). See also Additional file 3.

RNA-seq data mapping and variant detection

For all the samples, RNA-seq variant were detected using the snakemake [42] pipeline, available at this reference: [43]. Briefly, STAR v.2.5.2b was used for the read mapping on the Gallus_gallus-5.0 reference genome, after the multi-sample 2-pass mapping procedure, with a GTF file enriched in long non-coding genes that we recently published REF as input file for the

generation of genome indexes step as described in Muret et al. [6], Samples were analysed by tissue. FASTQ files were previously trimmed for Illumina adapter using TrimGalore version 0.4.5. Variant detection was done for each sample using the HaplotypeCaller function of GATK [23, 44, 45] 3.7.0 with option “--stand_call_conf 20.0”, “--min_base_quality_score 10” and “--min_mapping_quality_score 20”, generating one GVCF file per sample. The “GenotypeGVCFs” function was then used to combine these GVCF into one VCF per population. Biallelic SNP were then extracted using the SelectVariant function with option "--restrictAllelesTo BIALLELIC”. Variant were filtered using “VariantFiltration” with two of the three suggested filters, as we discuss in the Results and Discussion section: “QD < 2” and “FS > 30”.

DNA-seq read mapping and variant detection

The “GenotypeGVCFs” function was then used to combine these GVCF into one VCF for the 15 RpRm and one VCF for the 8 FLLL. Biallelic SNP were then extracted using the SelectVariant function with option "--restrictAllelesTo BIALLELIC”. Variant were filtered using “VariantFiltration” with all the recommended filters for DNA-seq: "FS > 60.0", "QD < 2.0", "MQ < 40.0", "MQRankSum < -12.5", "ReadPosRankSum < -8.0" and "SOR > 3.0".

Gene and exon expression quantification

Gene expression was quantified with RSEM [46] v.1.3.0, using the extended GTF file at the gene-level. To compute exon level expression, we used FeatureCount v1.6.2 with options -t “exon” and -g “exon_id”. We defined for each exon a metric called RpKb (Read per Kilobase) as the mean number of readsread mapped at the exon divided by its length in kilobases. To define an expression threshold, we compared the exon expressions to the expression of a set of randomly selected loci in the genome (see [article 1 de la présente thèse] for details). The expression of these regions were well below the expression of the exons. We set as an expression threshold for the exons an RpKb value of 0.5, corresponding the first quartile of expression in both RpRm and FLLL (see Additional file 4).

Variant functional predictions

Variant Effect Predictor (VEP) v92 was used for the effect prediction of the 9,949,072 SNP.

Homopolymers and exon-exon junction detection

Regions with 5 or more repeated nucleotides (homopolymers) and the regions spanning 5 bp of each extremity of a junction were detected using home-made scripts.

Multivariate analysis

The HAC was realized using the function "snpgdsHCluster" from the R [47] package SNPRelate v1.8.0 [48].

Bibliography

1. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5:621–8.

2. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369:1318–30.

3. Jehl F, Désert C, Klopp C, Brenet M, Rau A, Leroux S, et al. Chicken adaptive response to low energy diet: main role of the hypothalamic lipid metabolism revealed by a phenotypic and multi-tissue transcriptomic approach. BMC Genomics. 2019;20:1033.

4. Savary C, Kim A, Lespagnol A, Gandemer V, Pellier I, Andrieu C, et al. Depicting the genetic architecture of pediatric cancers through an integrative gene network approach. Sci Rep. 2020;10:1224.

5. Gondret F, Vincent A, Houée-Bigot M, Siegel A, Lagarrigue S, Causeur D, et al. A transcriptome multi-tissue analysis identifies biological pathways and genes associated with variations in feed efficiency of growing pigs. BMC Genomics. 2017;18:244.

6. Muret K, Klopp C, Wucher V, Esquerré D, Legeai F, Lecerf F, et al. Long noncoding RNA repertoire in chicken liver and adipose tissue. Genet Sel Evol. 2017;49:6.

7. Lagarrigue S, Martin L, Hormozdiari F, Roux P-F, Pan C, van Nas A, et al. Analysis of Allele-Specific Expression in Mouse Liver by RNA-Seq: A Comparison With *Cis* -eQTL Identified Using Genetic Linkage. Genetics. 2013;195:1157–66.

8. Roux P-F, Frésard L, Boutin M, Leroux S, Klopp C, Djari A, et al. The Extent of mRNA Editing Is Limited in Chicken Liver and Adipose, but Impacted by Tissular Context, Genotype, Age, and Feeding as Exemplified with a Conserved Edited Site in COG3. G3. 2016;6:321–35.

9. Piskol R, Ramaswami G, Li JB. Reliable Identification of Genomic Variants from RNA-Seq Data. The American Journal of Human Genetics. 2013;93:641–51.

10. Genome Aggregation Database Consortium, Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581:434–43.

11. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nature Reviews Genetics. 2014;15:121–32.

12. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Research. 2012;22:1775–89.

13. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. 2008;40:1413–5.

14. Peng Z, Cheng Y, Tan BC-M, Kang L, Tian Z, Zhu Y, et al. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. Nat Biotechnol. 2012;30:253–60.

15. Quinn EM, Cormican P, Kenny EM, Hill M, Anney R, Gill M, et al. Development of Strategies for SNP Detection in RNA-Seq Data: Application to Lymphoblastoid Cell Lines and Evaluation Using 1000 Genomes Data. PLoS ONE. 2013;8:e58815.

16. Tang X, Baheti S, Shameer K, Thompson KJ, Wills Q, Niu N, et al. The eSNV-detect: a computational system to identify expressed single nucleotide variants from transcriptome sequencing data. Nucleic Acids Research. 2014;42:e172–e172.

17. Wang C, Davila JI, Baheti S, Bhagwate AV, Wang X, Kocher J-PA, et al. RVboost: RNA-seq variants prioritization using a boosting method. Bioinformatics. 2014;30:3414–6.

18. Wolfien M, Rimmbach C, Schmitz U, Jung JJ, Krebs S, Steinhoff G, et al. TRAPLINE: a standardized and automated pipeline for RNA sequencing data analysis, evaluation and annotation. BMC Bioinformatics. 2016;17:21.

19. Oikkonen L, Lise S. Making the most of RNA-seq: Pre-processing sequencing data with Opossum for reliable SNP variant detection. Wellcome Open Res. 2017;2:6.

20. Cornwell M, Vangala M, Taing L, Herbert Z, Köster J, Li B, et al. VIPER: Visualization Pipeline for RNA-seq, a Snakemake workflow for efficient and complete RNA-seq analysis. BMC Bioinformatics. 2018;19:135.

21. Adetunji MO, Lamont SJ, Abasht B, Schmidt CJ. Variant analysis pipeline for accurate detection of genomic variants from transcriptome sequencing data. PLoS ONE. 2019;14:e0216838.

22. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.

23. Auwera GAV der, Carneiro MO, Hartl C, Poplin R, Angel G del, Levy-Moonshine A, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Current Protocols in Bioinformatics. 2013;43:11.10.1-11.10.33.

24. Guo Y, Zhao S, Sheng Q, Samuels DC, Shyr Y. The discrepancy among single nucleotide variants detected by DNA and RNA high throughput sequencing data. BMC Genomics. 2017;18:690.

25. Gallego Romero I, Pai AA, Tung J, Gilad Y. RNA-seq: impact of RNA degradation on transcript quantification. BMC Biol. 2014;12:42.

26. Porath HT, Carmi S, Levanon EY. A genome-wide map of hyper-edited RNA reveals numerous new sites. Nature Communications. 2014;5:4726.

27. Olofsson B, Bernardi G. The distribution of CR1, an Alu-like family of interspersed repeats, in the chicken genome. Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression. 1983;740:339–41.

28. Carmi S, Borukhov I, Levanon EY. Identification of Widespread Ultra-Edited Human RNAs. PLoS Genet. 2011;7:e1002317.

29. Kleinman CL, Adoue V, Majewski J. RNA editing of protein sequences: A rare event in human transcriptomes. RNA. 2012;18:1586–96.

30. Tan MH. Dynamic landscape and regulation of RNA editing in mammals. :27.

31. Lagarrigue S, Hormozdiari F, Martin LJ, Lecerf F, Hasin Y, Rau C, et al. Limited RNA Editing in Exons of Mouse Liver and Adipose. Genetics. 2013;193:1107–15.

32. Frésard L, Leroux S, Roux P-F, Klopp C, Fabre S, Esquerré D, et al. Genome-Wide Characterization of RNA Editing in Chicken Embryos Reveals Common Features among Vertebrates. PLoS ONE. 2015;10:e0126776.

33. Shafiei H, Bakhtiarizadeh MR, Salehi A. Large-scale potential RNA editing profiling in different adult chicken tissues. Animal Genetics. 2019;50:460–74.

34. Picardi E, D'Erchia AM, Lo Giudice C, Pesole G. REDIportal: a comprehensive database of A-to-I RNA editing events in humans. Nucleic Acids Res. 2017;45:D750–7.

35. Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Research. 2012;40:W452–7.

36. Ng PC. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Research. 2003;31:3812–4.

37. Le Béguec C, Wucher V, Lagoutte L, Cadieu E, Botherel N, Hédan B, et al. Characterisation and functional predictions of canine long non-coding RNAs. Sci Rep. 2018;8:13444.

38. Sherry ST, Ward, M.-H., Kholodov M., Baker J., Phan L., Smigielski E. M., et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Research. 2001;29:308–11.

39. Kranis A, Gheyas AA, Boschiero C, Turner F, Yu L, Smith S, et al. Development of a high density 600K SNP genotyping array for chicken. BMC Genomics. 2013;14:59.

40. Bordas A, Tixier-Boichard M, Merat P. Direct and correlated responses to divergent selection for residual food intake in Rhode island red laying hens. British Poultry Science. 1992;33:741–54.

41. Roux P-F, Boitard S, Blum Y, Parks B, Montagner A, Mouisel E, et al. Combined QTL and selective sweep mappings with coding SNP annotation and cis-eQTL analysis revealed PARK2 and JAG2 as new candidate genes for adiposity regulation. G3: Genes, Genomes, Genetics. 2015;5:517–529.

42. Koster J, Rahmann S. Snakemake--a scalable bioinformatics workflow engine. Bioinformatics. 2012;28:2520–2.

43. Snakemake/1000RNASeq_chicken/calling · master · bios4biol / workflows. GitLab. https://forgemia.inra.fr/bios4biol/workflows/tree/master/Snakemake/1000RNASeq_chicken/calling. Accessed 28 Aug 2019.

44. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. https://genome.cshlp.org/content/20/9/1297.long. Accessed 18 Sep 2020.

45. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics. 2011;43:491–8.

46. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC bioinformatics. 2011;12:323.

47. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2019. https://www.R-project.org/.

48. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics. 2012;28:3326–8.

3. Mise en place d'un pipeline d'analyse de l'expression allèle-spécifique (ASE) par RNA-seq avec phASER et ses déclinaisons et analyse de l'ASE dans des lignées F_1

a) contexte et objectifs

Le travail précédent nous a permis de détecter grâce à des données de RNA-seq des SNP fiables et des génotypes fiables, ce qui est, nous l'avons vu à différentes reprises, fondamental pour pouvoir réaliser une analyse d'expression allèle-spécifique (ASE). D'autres étapes sont cependant nécessaires pour pouvoir réaliser une telle analyse, et nous présentons dans la suite notre travail de mise en place d'un pipeline idoine, en utilisant des outils disponibles et en traitant les résultats qu'ils fournissent à l'aide de scripts *ad hoc*. Nous nous servirons de ce pipeline d'ASE dans le cadre de notre recherche de gènes candidats causaux de l'efficience alimentaire (*cf*. partie III – article 5 en préparation).

Nous avons vu en introduction l'importance de la régulation de l'expression des gènes dans la variation des caractères complexes, et en particulier des régulations « en *cis* ». Les *cis*-régulations sont l'ensemble des régulations agissant directement sur le gène régulé, sans passage par une molécule intermédiaire. La seule manière de déterminer si un gène est *cis*-régulé est d'observer pour ce gène une expression allèle-spécifique (ASE) dans au moins un tissu d'un individu. Ceci suppose que le variant régulateur, qui est par ailleurs inconnu, soit à l'état hétérozygote dans cet individu et qu'il existe au moins un SNP à l'état hétérozygote dans les transcrits du gène en question exprimés dans le tissu étudié. Ce(s) SNP hétérozygote(s) dans les transcrits permet alors de déduire l'origine chromosomique (c'est-à-dire l'allèle) de chaque transcrit. Autant que possible, le ou les SNP hétérozygote(s) dans les gènes devraient être en déséquilibre de liaison avec le variant régulateur. Il est ainsi possible d'évaluer par RNA-seq l'expression de chaque allèle en comptant le nombre de *reads* portant chaque allèle du SNP hétérozygote dans le gène, une différence significative entre ces deux comptages indiquant alors une régulation en *cis*. Il est important de bien noter qu'il faut, pour qu'un gène soit analysable, qu'il porte au moins un SNP hétérozygote, sans quoi il est impossible de se prononcer sur sa nature de gène *cis*-régulé.

Ce travail sera utilisé pour identifier des gènes candidats causaux responsables d'une part de la différence de l'EA (et des autres caractères associés comme par exemple le gras corporel) observée entre les 2 lignées divergentes R+ et R- (*cf*. partie III – article 5 en préparation). Nous

chercherons en particulier des gènes régulés en *cis* en analysant l'expression allèle-spécifique dans des individus F_1 (issus de croisements réciproques R+ × R-), qui porteront donc des chromosomes issus de chacun de leurs parents, dans des traces de sélection (détectées par ailleurs). Ce schéma expérimental permet en effet de maximiser l'hétérozygotie de ces individus, aussi bien pour les variants régulateurs (qui provoqueront l'ASE) que pour les SNP dans les régions exprimées (qui permettront de l'observer). Vu la complexité du caractère d'intérêt, plusieurs tissus ont été analysés, dans différentes conditions physiologiques et à différents âges. Ainsi, et en guise de préambule à la partie III (article 5, en préparation) consacrée à la recherche de gènes causaux de l'efficience alimentaire et des caractères associés, nous présenterons brièvement dans la suite quelques résultats liés à l'ASE dans ces lignées, notamment quelques ordres de grandeurs concernant le nombre de gènes analysables ou encore le *fold-change* entre chromosomes, informations qui sont encore rares dans la bibliographie. Nous donnerons également des informations liées à phASER, en particulier des ordres de grandeurs concernant le nombre de segments détectés par gène dans la population ou encore le nombre de SNP par segment.

Dans la littérature, on trouve quelques outils permettant d'analyser l'expression allèle-spécifique, après un pipeline bio-informatique adapté (voir plus loin). Citons en particulier ASEReadCounter[201] développé par Stéphane Castel (article publié en 2015), et un outil plus récent développé par le même auteur, phASER[202] (article publié en 2016). Le premier compte pour chaque SNP hétérozygote le nombre de *reads* portant chaque allèle. Le second, à savoir phASER, étudie non plus les SNP un par un, mais phase au préalable tous les SNP possibles à l'échelle du gène, et ce dans chaque individu, sur la base des *reads*. Il forme ainsi ce que nous appellerons dans la suite des « segments ». Une fois ces segments générés, phASER compte dans chaque individu le nombre de *reads* associés à chacun des deux haplotypes de chaque segment. Il propose finalement pour chaque gène un seul segment, qui servira à le représenter, étape détaillée plus longuement dans la suite. Cette approche permet donc d'étudier l'ASE à l'échelle, idéalement, du transcrit entier. En revanche, elle gomme l'éventuelle variation de comptage par SNP, qui pourrait être due à l'existence de transcrits alternatifs. Il s'agit ici de la même approche que pour l'évaluation de l'expression des gènes pour laquelle l'expression à l'échelle du transcrit n'est en général pas utilisée, faute de données et d'outils appropriés. Nous avons opté dans la présente thèse pour l'outil phASER, utilisé également par le consortium GTEx[99,206].

La Figure 6 résume le pipeline d'analyse de l'expression allèle-spécifique (ASE) par RNA-seq depuis le 1er alignement des fichiers FASTQ (données d'origine contenant les *reads* de RNA-seq en sortie du séquenceur) à l'obtention des fichiers d'ASE par échantillon (un fichier par individu et par tissu) issus de « phASER gene AE ». Les étapes préliminaires à l'étude de l'ASE, à savoir la détection dans les données RNA-seq de variants de type SNP (qui correspondent au travail réalisé dans l'article précédent, article 2) sont d'abord succinctement décrites (§ *b* – panel 0 de la Figure 6). L'analyse de l'ASE proprement dite à l'aide de l'outil phASER et de ses déclinaisons est développée, jusqu'à l'obtention des fichiers de résultats issus de ces outils (§ *c* – panels 1 à 4 de la Figure 6). Le traitement de ces fichiers par individu et par tissu regroupant les informations de tous les individus par tissu est ensuite détaillé (§ *d*).

b) étapes préliminaires : détection de SNP par RNA-seq

Le tout premier point du pipeline (« 0) Alignement des reads sur le génome », Figure 6) est dans son principe commun à l'ensemble des applications possibles du RNA-seq : on prélève sur n individus des échantillons de différents tissus. Ces échantillons suivent le protocole idoine pour le séquençage des ARN, et on finit par obtenir des fichiers FASTQ contenant les courtes séquences bien souvent pairées qui sont alignées sur le génome de référence de l'espèce étudiée, donnant autant de fichiers BAM (*Binary Alignment Map*) qui contiennent entre autres la séquence des *reads* et leurs positions d'alignement sur le génome de référence. Notons que le fond des boîtes figurant les fichiers FASTQ et BAM sont ici de couleur sombre pour les différencier dans la suite. Le point suivant, (« 1) Détection de SNP à l'aide de données de RNA-seq », Figure 6) résume succinctement la démarche présentée plus en détail dans l'article 2 (*cf.* plus haut). La détection des variants localisés dans les régions exprimées de chaque tissu de chaque individu se fait à l'aide de différents outils de la suite GATK[348] (en particulier l'outil « `HaplotypeCaller` »[349]), auquel on fournit les fichiers BAM déjà produits, et qui ont par ailleurs pu être utilisés dans d'autres applications, en général pour l'analyse de l'expression des gènes. Cela nous permet d'obtenir des fichiers GVCF (*Genomic Variant Call Format*), avec toujours un fichier par individu par tissu. Pour chaque tissu, ces fichiers sont combinés entre individus, fournissant à chaque fois un fichier VCF (*Variant Calling Format*) qui recense l'ensemble des variants détectés dans les régions exprimées du tissu en question.

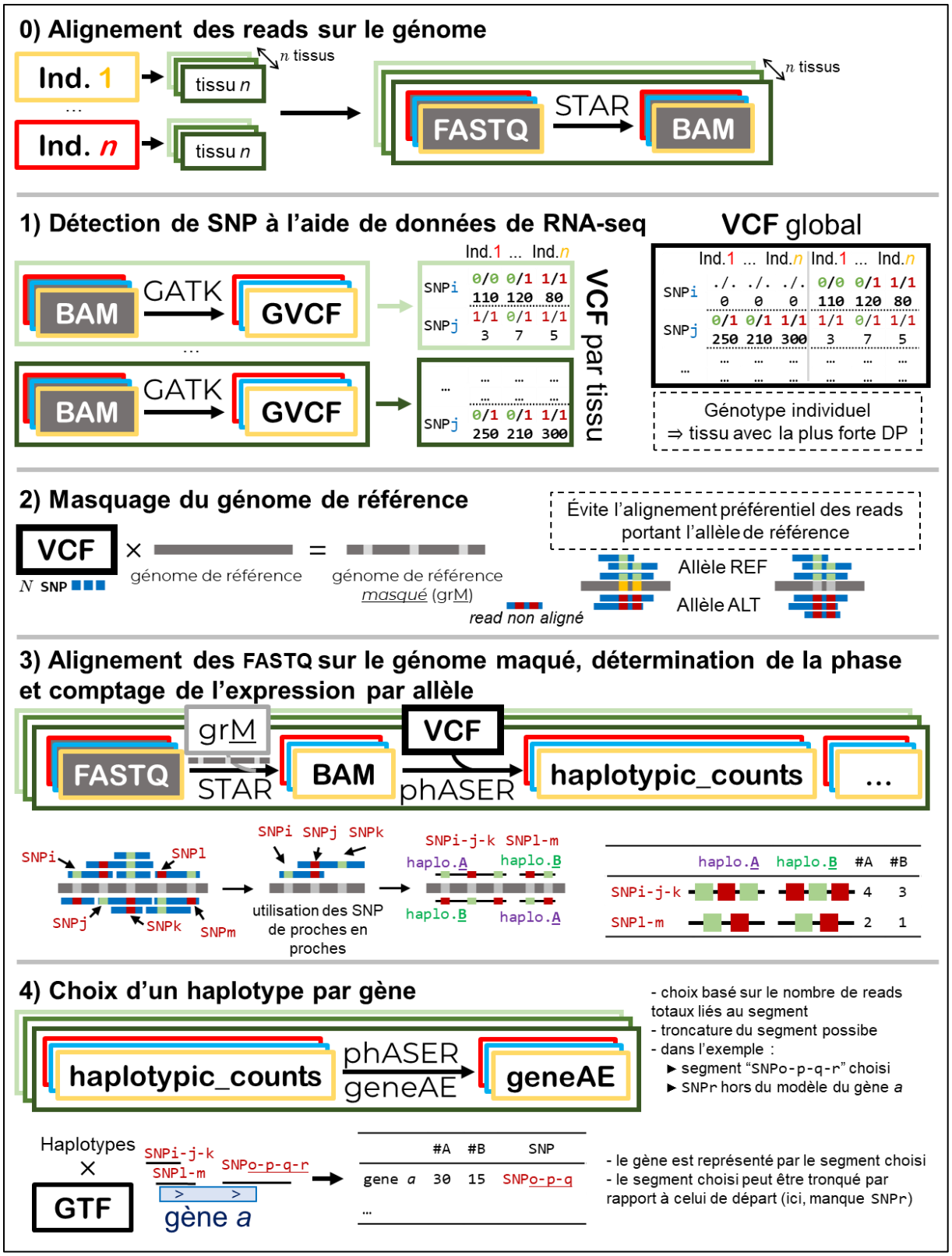


Figure 6 | Vue d'ensemble du *pipeline* d'analyse de l'expression allèle-spécifique (ASE) par RNA-seq. L'étape **0)** est l'étape classique débutant tout travail utilisant le RNA-seq, à savoir le prélèvement d'échantillons de tissu(s) (boîtes colorées en nuances de vert) sur différents individus (boîtes colorées en jaune, rouge et dans la suite, bleu, légendées « Ind. 1 » à « Ind. n »), puis, après les étapes de séquençages (non détaillées), l'alignement des *reads* sur le génome de référence. L'étape **1)** correspond au travail entrepris dans l'article 2 pour la détection de SNP par RNA-seq et la sélection de SNP fiables. *Suite de la légende en haut de la page suivante.*

DP : *depth*, soit le nombre de *reads* portant le variant L'utilisation de ces positions pour une analyse ASE implique de disposer d'un génome de référence masqué (grM) aux positions variables pour ne pas favoriser l'alignement des *reads* portant l'allèle de référence à ceux portant l'allèle alternatif. C'est l'objet de l'étape **2)**. Ceci fait, on passe à l'étape **3)**, qui consiste à aligner les *reads* générés dans l'étape **0)** sur ce génome masqué, et phASER utilise les *reads* pour phaser les variants entre eux, créant des segments qui ont chacun deux allèles par individu, appelés A et B. Plusieurs segments peuvent être créés pour un gène donné, et on voit que rien ne garantit que d'un gène à l'autre, l'allèle A vienne du même chromosome. grM : génome de référence masqué. Enfin, dans l'étape **4)**, on utilise une déclinaison de phASER pour assigner à chaque gène un segment. Le problème de l'éventuelle différence de chromosome d'origine des haplotypes « A » et « B » associés aux différents gènes subsiste, et le format du fichier de sortie nécessite des ajustements pour être utilisé. Ces dernières étapes sont présentées dans le texte.

Les critères présentés dans l'article 2 (SNP ayant (*i*) un génotype renseigné dans au moins 50% des individus d'au moins un tissu et (*ii*) dont au moins 20% des génotypes renseignés sont soutenus par au moins 5 reads), ont été appliqués aux SNP détectés par RNA-seq pour obtenir dans chaque tissu le génotype de chaque individu. Puisque plusieurs tissus étaient disponibles pour chaque individu, nous avons retenu, pour les SNP avec un génotype détecté dans plusieurs tissus, le génotype détecté dans le tissu où il est soutenu par le plus de *reads* afin de le fiabiliser autant que possible. Ainsi, on peut voir Figure 6, (« 1) Détection de SNP à l'aide de données de RNA-seq ») dans les VCF par tissu et dans le VCF global que le `SNPi` n'est détecté que dans le tissu vert clair (on prendra donc le génotype qui lui est associé dans ce tissu), et que le `SNPj` est détecté dans les deux tissus (vert clair et vert foncé), mais avec un nombre très différent de *reads* qui le soutiennent. D'ailleurs, le nombre de *reads* est si faible pour l'individu 1 du tissu vert clair que le génotype associé est `1/1` (homozygote pour l'allèle alternatif à l'allèle de référence), alors qu'il est `0/1` dans le tissu vert foncé (hétérozygote), sans doute parce que par hasard, les 3 *reads* utilisés pour déterminer le génotype dans le tissu vert clair contenaient tous l'allèle alternatif. En utilisant pour déterminer le génotype de l'individu à la position l'information soutenue par le plus de *reads*, on réduit le risque de ce genre d'erreurs.

c) utilisation de phASER pour l'étude de l'expression allèle-spécifique

La détection des SNP faite, passons à la première partie spécifique à l'ASE, à savoir le masquage du génome de référence (« 2) Masquage du génome de référence », Figure 6). Comme nous l'avions évoqué en introduction et comme nous l'illustrons ici, puisque l'ASE est basée sur l'utilisation de *reads* portant des SNP nécessairement hétérozygotes, une partie des *reads* porte l'allèle de référence (le même que celui du génome de référence donc, appelé REF dans la suite), quand l'autre partie porte un allèle alternatif (ALT dans la suite). Le problème est que lors de l'alignement, cet allèle alternatif sera considéré comme un « *mismatch* » par l'aligneur.

Selon les critères relatifs au nombre maximal de *mismatches* tolérés par *read* que lui fournit l'utilisateur, il est possible que certains *reads* portant des allèles ALT ne soient pas alignés. Compter le nombre de *reads* alignés portant chaque allèle risque alors de produire un résultat biaisé en faveur de l'allèle REF. Pour s'affranchir en partie de ce problème qui entrainerait pour environ 10% des SNP un biais de comptage > 5%[201,350], on peut « masquer » le génome de référence, c'est-à-dire remplacer les positions variables par des « *N* » dans le fichier qui contient le génome. Ainsi, à ces positions, les allèles REF et ALT seront considérés de la même façon et ne pénaliserons pas l'alignement des *reads*.

En utilisant le génome de référence ainsi masqué, il est ensuite nécessaire d'aligner une nouvelle fois les FASTQ contenant les *reads* de RNA-seq (« 3) Alignement des fastq sur le génome maqué, détermination de la phase et comptage de l'expression par allèle », Figure 6). Ce sont ces mêmes fichiers que ceux utilisés en « 0) », d'où la couleur sombre du fond des boîtes, mais cela génère de nouveaux fichiers BAM, d'où le fond clair pour ces boîtes. Ces fichiers BAM servent de fichiers d'entrée à phASER, accompagnés du VCF listant les SNP à étudier (ces mêmes SNP qui ont été masqué précédemment). À ce stade, dans chaque tissu de chaque individu, phASER utilise les *reads* qui se chevauchent pour phaser les variants détectés, comme nous l'illustrons dans la Figure 6. On voit par exemple que l'allèle REF (vert clair) du `SNPi` et l'allèle ALT (rouge foncé) du `SNPj` sont portés tous les deux par un même *read*. L'allèle ALT du `SNPj` est aussi porté par un autre *read* qui ne couvre pas la position du `SNPi` mais porte l'allèle REF du `SNPk`. Ainsi, même si les allèles des `SNPi` et `SNPk` ne sont jamais portés par un même *read*, on peut déduire, par l'intermédiaire du `SNPj`, les deux combinaisons alléliques dans l'individu : un haplotype REF – ALT – REF, issu d'un chromosome, appelé haplotype A par phASER et un haplotype ALT – REF – ALT, issu de l'autre chromosome, appelé haplotype B, pour le segment composé des `SNP i`, `j` et `k`. Il s'agit là d'un phasage utilisant l'information moléculaire, et non d'un phasage statistique, comme souvent pratiqué en génétique des populations dans des dispositifs pour lesquels on dispose des informations de génotypages et des liens de parentés. L'inconvénient de cette approche est que d'un groupe de SNP phasés à l'autre, rien ne garantit que les haplotypes appelés « A » et « B » par phASER viennent du même chromosome. Ainsi, dans notre exemple, l'haplotype A du segment de SNP `i`, `j` et `k` est issu du même chromosome que l'haplotype B du SNP `l` et `m`. Dans les faits, il nous semble que phASER appelle « haplotype A » l'haplotype pour lequel l'allèle du tout premier variant est l'allèle REF. Une fois les deux haplotypes déterminés, il ne reste plus qu'à leur assigner des

reads afin d'évaluer leurs expressions relatives. Pour ce faire, phASER compte simplement le nombre de *reads* qu'il peut assigner sans ambiguïté à chaque haplotype. En plus du fichier contenant les haplotypes et leurs comptages, phASER produit également un fichier contenant les comptages par SNP, et des fichiers contenant des informations sur les phases, pour chaque tissu de chaque individu.

L'étape précédente a donc permis de détecter un certain nombre de segments, dont plusieurs peuvent chevaucher un gène. L'étape suivante (« 4) Choix d'un haplotype par gène », Figure 6) consiste donc, comme son nom l'indique, à ne retenir qu'un seul segment par gène. Ceci fait, le gène et son expression seront d'une certaine façon représentés par ce segment. Il nous semble ici que l'haplotype choisi est celui qui est représenté par le plus de *reads*. C'est l'outil « phASER gene AE » qui est utilisé à cette fin : à partie d'un fichier d'annotation du génome (fichier GTF, *Gene Transfert Format*) et du fichier contenant les haplotypes générés par phASER, phASER gene AE sélectionne le meilleur segment et le tronque des variants situés hors du modèle du gène. Les raisons de ce dernier choix ne sont pas données par le créateur de phASER : peut-être l'outil est-il surtout utilisé avec des données d'espèces modèles, chez lesquelles les modèles de gènes sont bien connus, peut-être est-ce dû à l'observation qu'en RNA-seq, certains *reads* ont tendance à relier entre eux de nombreux modèles de gènes à cause d'un alignement incertain ? Toujours est-il que cela a tendance à exclure des segments des variants situés dans des extrémités 3' et 5' mal modélisées. En plus de la sélection du segment représentant chaque gène, phASER gene AE calcule également l'« *allelic fold-change* » définit comme le ratio entre les *reads* de l'haplotype A et ceux de l'haplotype B. À ce stade, nous disposons donc pour chaque tissu de chaque individu d'un fichier contenant les valeurs d'expression par haplotype de chaque gène qui a pu être analysé par phASER. C'est également à ce stade qu'est réalisé le test statistique permettant de déterminer si le ratio des comptages entre les deux haplotypes est significativement différent de 1 : on réalise donc un test binomial, en ne considérant que les gènes ayant plus de 10 *reads* associés à au moins un des deux allèles. Les *p-values* sont ensuite corrigées à l'échelle du génome entier en utilisant la méthode de Benjamini-Hochberg[351], avec un FDR (*False Discovery Rate*) de 0.05.

d) traitements post-hoc des résultats

Nous avons vu que pour chaque gène et chaque individu, phASER avait d'abord phasé les SNP présents dans les gènes, formant des segments, puis que phASER gene AE avait sélectionné un

de ces segments pour représenter le gène. Nous parlerons directement de gène dans la suite, pour alléger le discours.

Pour un gène donné, il s'agit maintenant de s'assurer que les haplotypes « A » et « B » du segment choisi viennent du même chromosome entre individus, avec la difficulté que d'un individu à l'autre, ce n'est pas forcément le même segment qui a été choisi pour représenter le gène. Pour nous simplifier la tâche, nous avons tiré parti du fait que les pedigrees des individus F_1 étudiés étaient connus et que les parents étaient séquencés par DNA-seq. En effet, nous nous sommes contentés d'assigner à chaque haplotype de chaque gène sa lignée d'origine (R+ ou R-), remplaçant ainsi le vocable « haplotypes A et B » par « haplotypes R+ et R- ». Dans le cas où ce type d'informations ne serait pas disponible, s'assurer de la cohérence des haplotypes « A » ou « B » pour un gène donné à travers les individus pourrait s'avérer nettement plus compliqué. Ceci fait, nous avons combiné les fichiers par individu en un grand fichier pour simplifier les analyses. Toutes ces étapes ont été réalisées à l'aide de scripts R[352] écrits à ces fins.

L'association de chaque haplotype à un chromosome parental consiste à déterminer l'origine parentale de chaque allèle composant les haplotypes, ce qui est possible tant qu'un allèle n'est présent que chez un seul des deux parents (cas où un parent est homozygote et l'autre hétérozygote ou bien où les deux parents sont homozygotes pour chaque allèle : parent 1 REF / ALT et parent 2 REF / REF pour le premier cas, parent 1 REF / REF et parent 2 ALT / ALT pour le second). Dans les cas où tous les variants d'un haplotype ne peuvent être assignés avec certitude, l'haplotype est assigné à la lignée à laquelle le maximum d'allèles a été assigné avec certitude.

e) exploration de l'expression allèle-spécifique dans le dispositif F_1 R+ × R- utilisé dans la partie III

Dispositif F_1 – Nous avons appliqué le pipeline d'analyse présenté ci-dessus à des données de RNA-seq multi-tissus d'animaux F_1 issus du croisement réciproque de parents R+ et R-, c'est-à-dire que la moitié d'entre eux avait un père R+ et une mère R-, alors que l'autre moitié avait un père R- et une mère R+. Des échantillons de quatre tissus – le tissu adipeux, le sang, l'hypothalamus et le foie – ont été prélevés sur 8 animaux, 4 mâles et 4 femelles, âgés de 35 semaines, à l'état nourri. Pour le foie, un organe majeur de l'homéostasie énergétique, dans lequel nous avons supposé l'existence d'évènements responsables d'une part de la variation de l' efficience alimentaire et des caractères associés (notamment la différence de gras corporel),

nous avons multiplié les conditions d'étude : des échantillons de foie ont ainsi été prélevés sur 8 animaux à 3 âges différents (9 jours, 10 semaines et 35 semaines) abattus à l'état nourri, et sur 8 animaux de 35 semaines abattus après un jeûne de 16h. Tous les animaux présentés ici avaient les mêmes répartitions de sexe et de parenté.

Résultats – Dans la mesure où l'approche adoptée par phASER repose sur le phasage des SNP présents dans les gènes, nous avons d'abord analysé le nombre de segments construits par phASER et le nombre de SNP composant ces segments. Toutes données confondues, nous avons dénombré en moyenne 3.5 segments par individu (médiane : 2, 1^{er} – $3^{è}$ quartile : 1 – 4). En ce qui concerne le nombre de SNP par segment, nous avons dénombré en moyenne 4.5 SNP par segment (médiane : 2, 1^{er} – $3^{è}$ quartile : 1 – 4 également). Après la sélection d'un segment par gène par phASER gene AE, les segments retenus sont composés de 7 SNP en moyenne (médiane : 3, 1^{er} – $3^{è}$ quartile : 1 – 9), pour tous les gènes et toutes données confondues. L'origine parentale de chacun des deux haplotypes a quant à elle pu être déterminée avec certitude pour 5 204 gènes en moyenne (médiane : 5 142, 1^{er} – $3^{è}$ quartile : 4 748 – 5 519), sur 7 987 gènes analysables en moyenne (médiane : 7 894, 1^{er} – $3^{è}$ quartile : 7 128 – 8 900).

Nous avons constaté que les segments associés aux gènes ASE étaient composés de significativement plus de SNP que ceux des gènes non-ASE pour tous les tissus, âges et conditions nutritionnelles analysées (test de Wilcoxon-Mann-Whitney, $p < 2.2\times10^{-16}$, échantillons de foie à 10 semaines : $p < 4\times10^{-6}$ et à 9 jours : $p < 3\times10^{-5}$; boîtes à moustaches blanches *versus* colorées, Figure 7A). Nous avons également observé que la somme des *reads* associés aux haplotypes A et B était significativement plus élevée (test de Wilcoxon-Mann-Whitney, $p < 2.2\times10^{-16}$) chez les gènes ASE versus les non ASE (Figure 7B). En observant le lien entre comptages totaux et *fold-change* d'expression entre les deux allèles pour les gènes ASE et non-ASE pour chaque individu, nous avons constaté que plus un gène avait un petit nombre de comptages totaux, plus le *fold-change* entre les deux allèles devait être fort pour être considéré comme significatif (Figure 7C, échantillons représentatifs de l'ensemble de échantillons). Ce phénomène vient de ce que le test binomial utilisé pour déterminer si le ratio entre les *reads* issus de chaque allèle est significativement différent de 0.05 est plus sensible à de petites variations lorsque les valeurs de comptages sont élevées. C'est ce qui est illustré sur la Figure 7D, à gauche et à droite. Cette figure représente la valeur de la *p-value* du test binomial en fonction des comptages de l'allèle B (axe des abscisses) et du *fold-change* entre haplotype A et haplotype B (axe des ordonnées, échelle $\log_2$). Le long d'une ligne, la valeur de la *p-value*

est constante. La ligne rouge représente une *p-value* de 0.05 et la bleue une *p-value* de 10^{-6}, soit de l'ordre de grandeur nécessaire pour être significatif à 0.05 après correction de Bonferonni pour 10 000 comparaisons, soit le nombre approximatif de gènes exprimés dans un tissu. Les valeurs pour les deux axes ont ici été choisis pour être similaires à celles du panel C. On constate par exemple que pour un comptage de l'haplotype B à 10^2 *reads*, le $\log_2$ *fold-change* entre A et B doit être d'environ 0.4 (fold-change de 1.3 environ) pour être significatif à $p = 0.05$, alors que pour un comptage de l'allèle B à 10^3, le $\log_2$ fold-change entre A et B doit être d'environ 0.1 (*fold-change* de 1.1 environ). La partie droite de ce panel représente les mêmes paramètres mais pour des valeurs de comptages B jusqu'au troisième quartile des valeurs observées dans nos données, et des valeurs de *fold-change* jusqu'au maximum observé. On peut conclure de ces observations que « l'insensibilité » du test binomial aux faibles valeur de *reads* ne permet pas de détecter des gènes *cis*-régulés parmi les gènes peu exprimés. Le nombre de gènes ASE présenté dans la suite est donc probablement sous-estimé, à cause (*i*) des gènes faiblement exprimés, comme nous venons de le voir, et (*ii*) des gènes « non-analysables » car sans variant hétérozygote dans leurs régions exprimées.

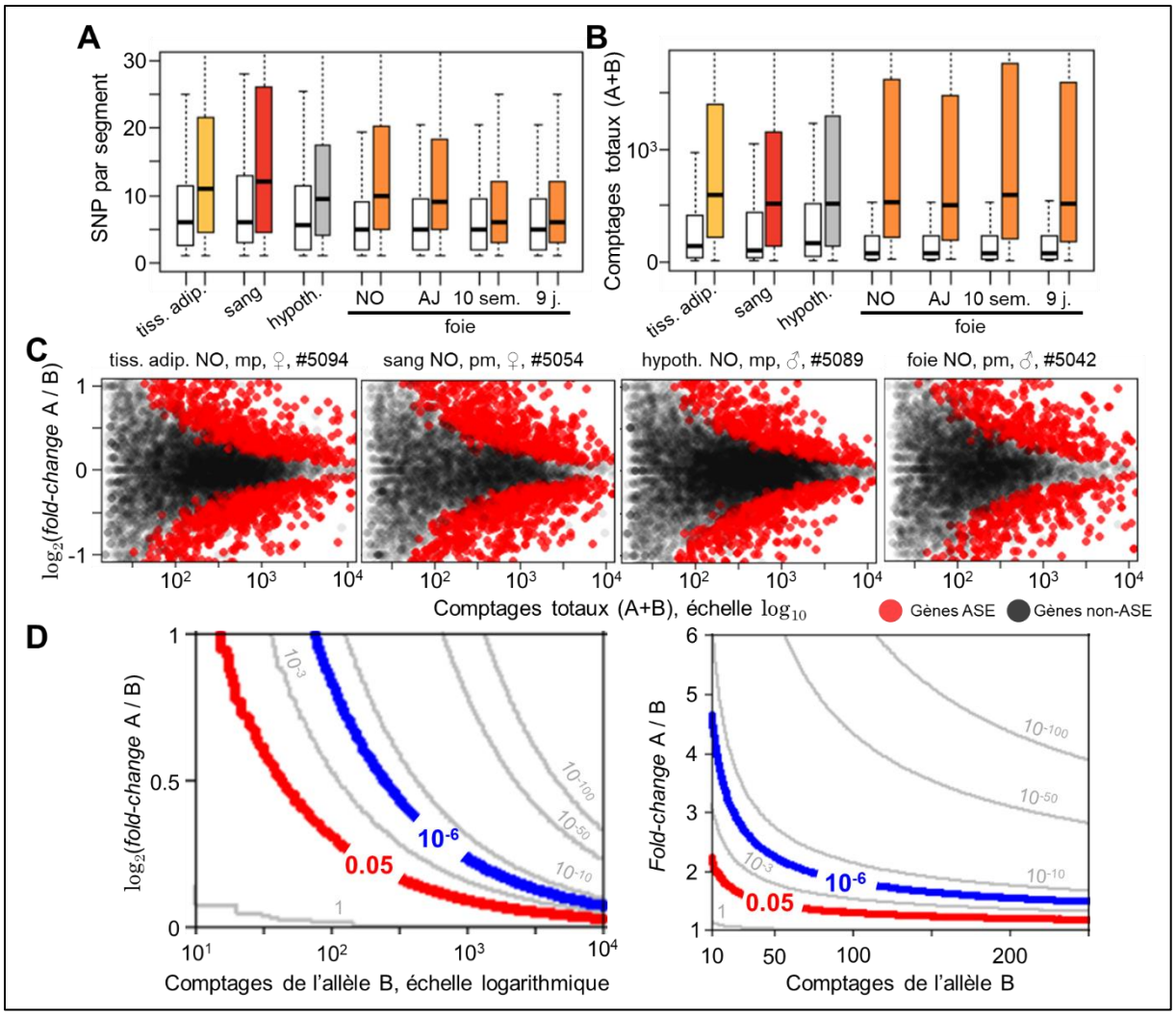


Figure 7 | Le test binomial pour l'ASE est biaisé en faveur des gènes plus exprimés. A. & **B**. Les gènes détectés comme ASE (boîtes à moustaches colorées) sont représentés par des segments composés de plus de SNP que les gènes non-ASE (boîtes à moustaches blanches). **A** Les segments en question sont couverts par plus de *reads*. **B** Comptage en *reads* par segments en échelle linéaire. **C**. Plus le nombre de comptages totaux associés aux segments est faible, plus le *fold-change* entre les deux allèles doit être important pour être significatif. **D**. *P-values* simulées de tests binomiaux en fonction du nombre de *reads* d'un allèle (ici, le B, axe des abscisses) et du *fold-change* entre A et B (axe des ordonnées), pour des valeurs du même ordre que sur les plots du panel C (gauche) et pour des valeurs équivalentes au troisième quartile des comptages d'un allèle et au maximum des *fold-change* plus larges (droite). En rouge, $p = 0.05$, en bleu, $p = 1\times10^{-6}$, soit l'ordre de grandeur du seuil de significativité à 0.05 pour 10 000 tests, après correction de Bonferroni.

Nous avons ensuite analysé (voir Tableau 2) le nombre de gènes présentant une ASE dans notre dispositif, en déclinant les résultats entre PCG et LNC. Sur les 18 000 gènes environ exprimés (TPM ≥ 0.1) en moyenne dans les tissus, âges et conditions physiologiques, 70% étaient des PCG et 30% des LNC. Parmi ces gènes, 43% en moyenne étaient analysables : 48% des PCG exprimés et 32% des LNC exprimés (7 820 gènes analysables en moyenne, dont 6 037 PCG et 1 742 LNC). Ces proportions sont cohérentes avec celles calculées dans l'article 2 dans le cadre de l'application consistant à évaluer le potentiel pour une analyse ASE dans les populations

étudiées. Pour les R+ et R-, nous montrions en effet que 45% des PCG exprimés à TPM ≥ 0.1 étaient analysables, et 23% des LNC.

Nous présentons ensuite les résultats de deux seuils différents pour considérer un gène comme étant ASE. Le premier seuil consiste à déclarer un gène comme ASE s'il est détecté comme significativement ASE ($p_{FDR} \leq 0.05$) dans au moins 2 individus, le second seuil, plus drastique, est de le déclarer ASE s'il est détecté comme significativement ASE dans au moins 50% des individus, ce qui correspond à 4 individus (3 dans l'hypothalamus). Quel que soit le seuil choisi, nous observons des proportions de gènes ASE par rapport aux gènes analysables assez similaires entre les tissus, tissu adipeux, hypothalamus et foies (à tous âges et conditions physiologiques) : avec le premier seuil, environ 24% en moyenne des gènes analysables sont ASE, aussi bien pour les PCG que les LNC. Ces résultats sont assez similaires à ceux trouvés dans la littérature (voir aussi l'introduction de la présente thèse), par exemple ceux obtenus par Zhuo *et al.*[125] en 2017 dans des foies d'embryon de poulets (15.4% des gènes analysables), par Lagarrigue *et al.*[111] (2013) dans des foies chez la souris (14.6% des gènes analysables), Edsgärd *et al.*[133] (2016) dans des globules blancs humains (13.6%) ou encore par le consortium GTEx[99] chez l'humain également, avec 22% et 26% de gènes *cis*-régulés par rapport aux gènes exprimés dans l'hypothalamus et le foie respectivement, et en ayant analysé 170 et 208 échantillons respectivement. En revanche, ce pourcentage passe à 53% dans le tissu adipeux, avec 469 échantillons analysés. Avec le second seuil, environ 10% en moyenne des gènes analysables sont ASE, avec 10% des PCG analysables, et 8% des LNC analysables. Le sang en revanche se démarque des autres tissus. Quel que soit le seuil, les proportions de gènes ASE dans le sang sont deux fois plus élevées que dans les autres tissus. Avec le premier seuil, 43% des gènes analysables sont ASE, et 50% des LNC analysables, et avec le second seuil 20% des gènes analysables, aussi bien PCG que LNC. Ces résultats ne sont pas retrouvés dans d'autres études ASE portant également sur le sang chez d'autres espèces. Par exemple l'étude de de Maroilley *et al.*[132] (2017) chez le porc, trouve que 11% des gènes analysables dans le sang sont ASE, en utilisant cependant un critère plus strict que le nôtre, puisqu'un gène y était considéré comme ASE s'il contenant au moins un SNP hétérozygote et ASE parmi au moins un tiers des animaux portant ce SNP. Le consortium GTEx[99] trouve pour sa part que 63% des gènes exprimés dans le sang total humain sont *cis*-régulés, en analysant 670 échantillons.

Tableau 2 | Vue générale de l'étendue de l'ASE dans les différents tissus, âges et stades physiologiques du dispositif F_1.

	35 semaines, nourris								35 sem., à jeun		10 sem., nourris		9 jours, nourris	
	Tissu adipeux		Sang		Hypoth.		Foie		Foie		Foie		Foie	
Exprimés	20112		16602		22257		16744		16742		17378		17025	
PCG	13469		10779		14553		12048		12085		12382		12311	
LNC	6485		5678		7528		4586		4555		4881		4600	
Analysables	8 972	(44.6%)	7083	(42.7%)	8766	(39.4%)	7467	(44.6%)	7379	(44.1%)	7535	(43.4%)	7541	(44.3%)
PCG	6622	(49.2%)	5496	(51.0%)	6602	(45.4%)	5858	(48.6%)	5804	(48.0%)	5943	(48.0%)	5932	(48.2%)
LNC	2316	(35.7%)	1532	(27.0%)	2108	(28.0%)	1565	(34.1%)	1559	(34.2%)	1583	(32.4%)	1528	(33.2%)
ASE ≥ 2 échant.	2434	(27.1%)	3007	(42.5%)	1970	(22.5%)	1872	(25.1%)	1794	(24.3%)	1816	(24.1%)	1739	(23.1%)
PCG	1934	(29.2%)	2237	(40.7%)	1419	(21.5%)	1466	(25.0%)	1408	(24.3%)	1428	(24.0%)	1363	(23.0%)
LNC	480	(20.7%)	747	(48.8%)	534	(25.3%)	395	(25.2%)	372	(23.9%)	376	(23.8%)	365	(23.9%)
$\log_2$ allelic fold-change	≥ 0.1	≥ 1 TPM	≥ 0.1	≥ 1 TPM	≥ 0.1	≥ 1 TPM	≥ 0.1	≥ 1 TPM	≥ 0.1	≥ 1 TPM	≥ 0.1	≥ 1 TPM	≥ 0.1	≥ 1 TPM
PCG	0.55	0.55	0.61	0.60	0.53	0.52	0.60	0.60	0.60	0.59	0.60	0.59	0.61	0.60
LNC	1.05	1.06	1.19	1.13	1.17	1.17	1.03	1.00	1.07	1.07	1.10	1.07	1.02	1.00
ASE ≥ 50% échant.	1023	(11.4%)	1366	(19.3%)	609	(6.9%)	771	(10.3%)	711	(9.6%)	761	(10.1%)	727	(9.6%)
PCG	822	(12.4%)	1034	(18.8%)	420	(6.4%)	607	(10.4%)	568	(9.8%)	612	(10.3%)	588	(9.9%)
LNC	195	(8.4%)	322	(21.0%)	184	(8.7%)	159	(10.2%)	139	(8.9%)	144	(9.1%)	136	(8.9%)

Les gènes sont considérés comme exprimés pour TPM ≥ 0.1 et analysables lorsqu'ils présentent au moins un SNP hétérozygote. Les valeurs pour les gènes analysables et les gènes ASE sont les médianes des individus. Le pourcentage de gènes analysables est exprimé en fonction des gènes exprimés, les pourcentages des gènes ASE sont exprimés en fonction de la médiane des gènes analysables. Les valeurs de $\log_2$ *allelic fold-changes* sont les médianes des individus, données ici pour les gènes ASE dans ≥ 2 échantillons et pour les gènes exprimés à ≥ 0.1 TPM et ≥ 1 TPM.

Pour finir, la répartition des valeurs absolues des *allelic fold-changes* ($\log_2$aFC, le $\log_2$-ratio des comptages associés à chaque allèle) pour les gènes ASE dans au moins deux individus montre dans tous les tissus des $\log_2$aFC de l'ordre de 0.6 pour les PCG (soit un *fold-change* de 1.5 environ) et de l'ordre de 1.10 pour les LNC (soit un *fold-change* de 2.14 environ). Il est intéressant de constater que même pour une valeur minimale de TPM ≥ 1 pour considérer un gène comme exprimé, PCG et LNC présentent tout de même des valeurs médianes de $\log_2$aFC assez différentes, du même ordre que celles obtenues avec le seuil de TPM ≥ 0.1. Cela signifie que le $\log_2$aFC plus élevé pour les LNC n'est pas dû uniquement au fait que ces gènes sont moins exprimés, et ont donc moins de *reads* associés, provoquant des différences de proportions plus importantes. Des résultats similaires sont obtenus pour les gènes ASE dans ≥ 50% des échantillons (*résultats non montrés*). Le consortium GTEx a également noté que les *cis*-eQTL affectant les LNC avaient des effets plus forts (au sens du $\log_2$ *allelic fold-change*) que ceux affectant les PCG, avec des médianes de valeurs absolues de $\log_2$ *allelic fold-change* de 0.805 pour les LNC contre 0.579 pour les PCG[273].

Conclusion – Comme nous l'avons vu dans l'article précédent, au moins 50% des PCG apparaissent comme analysables (c'est-à-dire portant au moins un SNP hétérozygote dans leurs régions transcrites), et environ 30% des LNC sans doute en raison de leur expression plus faible

qui n'est pas compensée par leur densité plus importante en SNP (rappelons en effet que la séquence primaire des LNC subie une pression de sélection moins forte que celle des PCG).
En moyenne, un quart des gènes analysables sont ASE, proportion par ailleurs stable entre tissus ou conditions physiologiques pour un tissu donné. En revanche, on peut s'interroger sur le doublement de cette proportion pour le sang par rapport aux autres tissus étudiés, qui n'est pas retrouvé dans la littérature. À notre connaissance il n'y a pas d'études réellement comparables, dans lesquelles le sang et au moins un autre tissu auraient été étudiés. Cette observation pourrait être due à un artéfact technique ou un phénomène biologique spécifique à ce tissu, par exemple lié au fait que le sang des non-mammifères contient des globules rouges nucléés, pourvus de mitochondries[353] et à priori transcriptionnellement actifs[354], mais il est difficile de conclure à ce stade de l'observation. Dans les autres tissus, les 25% des gènes ASE parmi les gènes analysables correspondent à l'ordre de grandeur retrouvé dans des études faites sur différentes espèces et tissus, indiquant au moins un quart de régulations de type *cis* dans les tissus, et avec des *fold-changes* plus importants pour les LNC que les PCG, ce qui a été observé par ailleurs[273,355] sans pour autant les expliquer.

II – Étude des gènes et processus biologiques impliqués dans la différence d'efficience alimentaire ainsi que dans l'adaptation des poules à des variations de régimes

Un des objectifs de la présente thèse était de mieux comprendre la composante génétique de l'efficience alimentaire (EA). Nous avons d'abord analysé les variations transcriptomiques entre les R+ et les R- dans différents tissus choisis pour leur rôle potentiel dans les différences de RFI entre les deux lignées (§ 1). Nous avons profité de l'annotation étendue du génome de la poule, enrichie en LNC, générée dans la partie précédente (voir partie I) pour chercher des LNC potentiellement impliqués dans les différences entre les lignées. Dans le § 1 ci-après, la valorisation des résultats obtenus avec l'analyse des PCG se fera séparément de celle des résultats obtenus pour les LNC. Notons que les listes de gènes différentiellement exprimés entre lignées seront exploitées dans la partie III dont un objectif est d'identifier des gènes causaux à l'EA de par une *cis*-régulation.

Dans cette partie II, nous avons ensuite cherché une éventuelle interaction entre l'EA et le contenu énergétique de l'aliment en analysant les conséquences d'un aliment hypo-énergétique sur les transcriptomes tissulaires d'animaux de ces deux lignées (§ 2).

1. Identification de gènes et voies métaboliques associées avec la variation d'efficience alimentaire par des analyses multi-omics dans le foie et le tissu adipeux (article 3)

a) contexte et objectifs

L'amélioration de l'efficience alimentaire demeure un défi important pour assurer la rentabilité de filières et réduire l'impact environnemental de l'élevage. En effet, l'aliment destiné aux animaux représente 60% des coûts de production, et sa production est responsable d'une part importante des impacts environnementaux de l'élevage. Quoique la poule fasse partie des animaux d'élevage les plus efficients, une meilleure compréhension des mécanismes sous-jacents à ce caractère, et la poursuite de l'amélioration de l'EA n'en restent pas moins des objectifs clefs. Pour mieux comprendre les mécanismes impliqués dans l'efficience alimentaire, nous avons étudié les différences d'expressions géniques et de composition du lipidome dans le foie et le tissu adipeux des deux lignées divergentes pour l'efficience alimentaire résiduelle.

Comme nous l'avons dit, en plus de leur différence marquée d'efficience alimentaire résiduelle, ces deux lignées présentent également une différence marquée d'adiposité corporelle, la lignée R- (efficiente) ayant une masse de tissu adipeux abdominal environ deux fois plus élevée que la lignée R+ (inefficiente). Nous avons donc fait l'hypothèse que le foie et le tissu adipeux pourraient être à l'origine de certaines de ces différences phénotypiques, et avons donc étudié ces deux tissus par transcriptomique. En effet, ces deux tissus ont des rôles clefs dans l'homéostasie énergétique chez les oiseaux : le foie est le tissu de synthèse des lipides, notamment les acides-gras et le cholestérol, et le tissu adipeux est le lieu de stockage et de mobilisation de ces acides-gras, qui composent par ailleurs la majorité des réserves énergétiques des animaux. Différentes études se sont intéressées aux différences transcriptomiques entre des animaux efficients et inefficients chez des espèces d'élevage, y compris le poulet de chair, mais aucune ne s'est intéressée aux poules pondeuses, dont la physiologie et le métabolisme diffèrent fortement de ceux des poulets de chairs. Les résultats obtenus chez ces derniers ne se transposent donc pas chez les poules pondeuses, justifiant ainsi la présente étude.

b) matériels et démarches

Nous avons utilisé des données de RNA-seq, de lipidome et de métabolome collectées sur trois groupes d'animaux R+ et R- : groupe contrôle (CT), un groupe nourri avec un régime hypo-énergétique (LE, utilisé dans l'article 4 ci-après, avec le groupe CT), un groupe exposé à un stress thermique (HS, données non étudiées dans la présente thèse). Les données de RNA-seq multi-tissus proviennent de ces trois groupes pour l'hypothalamus (non étudié dans l'article mais dont nous parlerons dans le § *f*) et le foie (2 lignées × 24 poules, n = 8), du premier et du dernier groupe pour le tissu adipeux (2 lignées × 16 poules, n = 8), et enfin, des trois groupes pour le sang (2 lignées × 24 poules, n = 8, non étudié dans l'article mais dont nous parlerons dans le § *f*), en plus de données prélevées sur les animaux du groupe HS avant le début du stress (2 lignées × 8 poules). Les données de lipidomique du tissu adipeux utilisées ont été collectées sur des individus des groupes CT, pour un total de 24 animaux (2 lignées × 12 poules), et celles du foie ont été collectées sur des individus des groupes CT et HS, pour un total de 48 animaux (2 lignées × 2 groupes × 12 poules). Enfin, les données de métabolomique ont été collectées sur du foie d'individus des groupes CT, LE et HS, pour un total de 72 animaux (2 lignées × 3 groupes × 12 poules). Les données de transcriptomique et de lipidomique ont d'abord été analysées individuellement, en utilisant le package R « edgeR » pour les premières et par analyse de variances pour les secondes, en utilisant à chaque fois un modèle statistique tenant compte de la lignée, de la condition (CT, LE, HS), et de l'interaction entre ces termes. Une *p*-

value ajustée pour les tests multiples (méthode de Benjamini-Hochberg[351]) de 0.05 a été utilisé pour définir une variable significativement différente pour le facteur testé. Aucune interaction n'a d'ailleurs été détectée ($p_{FDR} \leq 0.05$). Ensuite, eu égard au grand nombre de gènes différentiellement exprimés détectés, et afin d'étudier conjointement les 3 jeux de données, nous avons réalisé une Analyse Factorielle Multiple (AFM) sur 26 animaux (13 R+ et 13 R-) pour lesquels les trois types de données étaient disponibles. L'AFM permet d'étudier les trois jeux de données en s'assurant qu'ils jouent tous un rôle équivalent dans l'analyse, et que le jeu de données transcriptomique (environ 20 000 variables) « n'écrase » pas les autres (100 variables au plus). Cette analyse nous a permis de mettre en évidence un petit jeu de gènes, de lipides et de métabolites participant le plus à la séparation des animaux des deux lignées. Nous avons alors cherché au sein des gènes des modules de co-expression, qui suggèrent un lien biologique entre les gènes, à l'aide de WGCNA. Nous avons enfin tenté de lier ces modules de gènes aux lipides et métabolites mis en évidence.

c) résultats

Parmi les 6 004 gènes différentiellement exprimés dans le foie dont 3 781 gènes codant des protéines, nous avons tout d'abord observé la sur-expression, dans la lignée R- (efficiente) de gènes impliqués dans la glycolyse, la lipogenèse *de novo*, le transport des acides gras, la synthèse des triglycérides, le stockage des lipides et leur exportation. Concomitamment, nous avons observé la sous-expression de gènes liés au métabolisme de différents acides-aminés, mais également à l'entrée du cholestérol dans les hépatocytes, son transport et la synthèse d'acides biliaires (qui se fait à partir du cholestérol, voir aussi Figure 1 de l'article). Ces résultats au niveau transcriptomique étaient cohérents avec ceux observés au niveau lipidomique : les quantités d'acides-gras et de triglycérides étaient environ 4 à 5 fois plus élevées dans le foie des R-, et la part d'acides gras mono-insaturés (issus de la lipogenèse *de novo*) y était également plus élevée que chez les R+ (Figure 3 de l'article).

Dans le tissu adipeux (Figure 2 de l'article), nous avons observé parmi les 5 032 gènes différentiellement exprimés dont 3 177 codant une protéine, une sur-expression dans les R- de gènes codant pour des composantes du complexe I de la chaine respiratoire mitochondriale, de gènes impliqués dans la matrice extracellulaire et dans l'inflammation. En revanche, les gènes sous-exprimés dans le tissu adipeux des R- n'ont pas révélé d'enrichissement significatif en termes fonctionnels. Si les deux lignées présentaient bien un fort contraste en termes de masse de tissu adipeux (deux fois plus pour les R-), au niveau lipidomique (Figure 3 de l'article), les quantités des différents lipides mesurés ne différaient pas entre les deux lignées, à l'exception

des proportions d'acides gras qui étaient similaires à celles observés dans le foie, pour les deux lignées.

L'analyse intégrative du transcriptome, du lipidome et du métabolome hépatique par AFM a permis de mettre en évidence 1201 gènes, 20 lipides et 23 métabolites qui séparaient particulièrement les R+ des R-. Parmi les gènes, nous avons mis en évidence grâce à WGCNA 4 modules de gènes fortement corrélés chez les R- et un seul seulement chez les R+. Les modules R- étaient enrichis en gènes liés au transport des vésicules du Golgi, à la glycolyse ou la néoglucogenèse, à la voie de signalisation des PPAR et au métabolisme des acides gras, alors que le module R+ ne présentait aucun terme fonctionnel enrichi. Parmi le top 10 des gènes codant des protéines les plus connectés aux autres gènes du module, nous avons observé beaucoup de gènes codant des enzymes, et très peu de gènes codant des facteurs de transcription. Notons tout de même la présence de *THRSP* (alias *SPOT14*) dans le module enrichi en gènes associés au terme « *fatty acid metabolism* », un gène qui semble être un régulateur de l'expression de gènes codant des enzymes clefs de la lipogenèse.

d) discussion et conclusion

Ce travail nous a permis de clairement montrer les différences liées au métabolisme des lipides entre les R+ et les R- aux niveaux transcriptomiques et lipidomiques, dans deux tissus clefs de ce métabolisme. En particulier, la différence de masse de tissu adipeux abdominal entre les deux lignées s'explique, d'abord, par la synthèse massive d'acides gras qui a lieu dans le foie des animaux R-, synthèse qui s'accompagne d'une forte activité d'exportation des lipides mais également par une stéatose hépatique. Les régulations sous-jacentes à cette synthèse *de novo* apparaissent complexes. En effet, ni *NR1H3* ni *SREBF1*, deux gènes majeurs dans la régulation de la lipogenèse *de novo* n'étaient différentiellement exprimés. En revanche, *THRSP*, un apparent régulateur de l'expression de gènes codant des enzymes clefs de la lipogenèse était différentiellement exprimé, ainsi que *FOXO1*, qui est impliqué dans la voie de signalisation de l'insuline et qui contribue à la stéatose hépatique dans des contextes d'insulino-résistance. Les R- présentent différents signes d'insulino-résistance, observés dans d'autre travaux : glycémie à jeun plus élevée que les R+, insulinémie également plus élevée, à jeun comme nourris.

La différence de masse de tissu adipeux abdominal entre les deux lignées s'explique ensuite par un stockage tout aussi massif de ces acides gras dans le tissu adipeux de la lignées R-. Ce stockage s'accompagne d'éléments évoquant une « surcharge mitochondriale », qui se traduit par l'activation du complexe I de la chaine respiratoire, provoquant la génération d'espèces

réactives de l'oxygène (*Reactive Oxygen Species*, ROS), et par suite, une inflammation, pouvant expliquer que des marqueurs inflammatoires et des gènes codant pour des protéines antioxydantes soient surexprimés dans ce tissu. De plus, le tissu adipeux présente des signes de remodelage, phénomène grâce auquel il accommode l'arrivée des lipides, notamment la surexpression de gènes liés à la matrice extracellulaire.

Enfin, l'étude des mécanismes impliqués dans le dialogue entre les deux tissus nous a fait soupçonner un possible dysfonctionnement des récepteurs hépatiques à l'adiponectine (qui est produite par le tissu adipeux). En effet, l'adiponectine, significativement surexprimée par le tissu adipeux des R-, est censée induire une diminution de la lipogenèse dans le foie. Cependant, elle agit à travers deux récepteurs dont l'un est sur-exprimé dans le foie des R-. Or, nous n'observons ni différence de quantité de ce lipide entre les foies des deux lignées, ni différence d'expression des gènes hépatiques cibles de la régulation de l'adiponectine.

Il apparait à travers ce travail que les R- présentent un phénotype de stéatose hépatique, un phénotype qui est de plus en plus étudié aujourd'hui puisqu'il est lié aux maladies métaboliques humaines, de type obésité ou diabète de type II. Ces données transcriptomiques hépatiques seront donc probablement ré-analysées sous cet angle, en lien avec des modèles similaires de souris, pour mieux comprendre les facteurs de la stéatose hépatique.

e) article, en cours de relecture par les co-auteurs

Cet article est en cours de relecture par ses co-auteurs : **Frédéric Jehl**, Colette Désert, Christophe Klopp, Andrea Rau, Morgane Boutin, Laetitia Lagoutte, Yuna Blum, Diane Esquerré, David Gourichon, Thierry Burlot, Yuliaxis Ramayo-Caldas, Anne Collin, Sandrine Lagarrigue, Tatiana Zerjal. Genes and biological pathways associated with layers variation in feed efficiency identified by liver and adipose tissue multi-omic analyses. Soumission prévue à *BMC Genomics* fin 2020.

Il est reproduit ci-après.

Genes and biological pathways associated with layers variation in feed efficiency identified by liver and adipose tissue multi-omic analyses.

Frédéric Jehl, Colette Désert, Christophe Klopp, Andrea Rau, Morgane Boutin, Laetitia Lagoutte, Yuna Blum, Diane Esquerré, David Gourichon, Thierry Burlot, Yuliaxis Ramayo-Caldas, Anne Collin, Sandrine Lagarrigue*, Tatiana Zerjal*

*Corresponding authors

Background

In the mono-gastric sector, feed represent 60 to 70% of the production costs and efforts have been invested to improve feed efficiency to increase production profitability and the environmental sustainability through reduced use of inputs. Although chicken is among the most efficient livestock, the comprehension of the underlying mechanisms of feed efficiency (FE) and the continuation of its improvement remains a key challenge. Feed efficiency is a complex trait often summarized by the residual feed intake (RFI) index that is a statistically-built index obtained as the difference between the observed feed intake and a predicted feed intake, estimated based on the animal maintenance requirements and the production activity of interest [1]. variation in RFI reflects differences in the efficiency of animals to use the ingested feed to maintain physiological functions and production capacity.

In order to study this complex trait in layer chicken, we used two lines divergently selected from a base Rhode Island Red egg-laying chicken population since 1976 on the residual feed intake (RFI) at equal body weight and egg production [2,3]. The two lines obtained are the R+ line selected for high RFI value and representing the low efficient line and the R- line selected for low RFI values and representing the high efficient line. After more than 40 generations of RFI divergent selection, the R+ hens consume on average 70% more feed than their counterpart R- hens, have higher diet-induced thermogenesis [4–6] but a reduced body fat content [7]. Although these lines have been widely studies from a phenotypic viewpoint, little is known on the underlying genetic mechanisms contributing to the divergence observed. In this study we investigate the genetic basis of feed efficiency by proposing an integrated analysis combining RNA-seq transcriptome, lipidome and metabolome data from liver and adipose tissue. The adipose tissue is the tissue where fatty acids, that are the main form of energy, are stored. The liver is a key organ for lipogenesis in birds [8], in addition to many other physiological processes such as oxidation, secretion and detoxification. Different studies have investigated

the transcriptomic differences between feed efficient and feed inefficient farm animals such as pigs [9–16], beefs [17–24] or broilers [25–30]. However, until now, no such work has been undertaken in layers. The physiology and the metabolism of layer chicken differ from that of broilers. Indeed, from hatching, significant differences between the two chicken types are observed in feed intake, growth rate, development of muscles and adipose tissue and in adults, in laying rates [31], therefore results obtained from broilers are not necessarily transposable to layers. In the present study we have first analyzed the differential expression of hepatic and adipose tissue genes in adult hens from the R+ and R- divergent lines, to improve our understanding of the genetic mechanisms involved in feed efficiency in layers. Since the hepatic differentially expressed genes were associated to terms related to the metabolism (more than 300 associated genes in total among the over- and under-expressed genes), we analyzed the liver transcriptomic data conjointly with liver lipidomic and metabolomics data, in order to improve our capacity to identify key pathways separating the two lines before searching for modules of co-expressed genes and study their correlations with the lipids and metabolites.

Results

Functional characterization of hepatic transcriptome differences between the R+ and R- lines

Term	p_{FDR}	Nb Genes
A. Over-expressed in R- versus R+		
Protein processing in endoplasmic reticulum	9.67×10^{-06}	40
Ribosome	1.84×10^{-05}	34
Metabolic pathways	1.60×10^{-03}	152
Galactose metabolism	5.90×10^{-03}	12
Amino sugar and nucleotide sugar metabolism	5.90×10^{-03}	15
B. Under-expressed in R- versus R+		
Metabolic pathways	3.37×10^{-15}	195
Tryptophan metabolism	3.02×10^{-06}	19
Carbon metabolism	7.12×10^{-06}	31
Glycine, serine and threonine metabolism	9.35×10^{-05}	16
Valine, leucine and isoleucine degradation	6.60×10^{-04}	16
Glyoxylate and dicarboxylate metabolism	9.50×10^{-04}	12
Biosynthesis of amino acids	1.12×10^{-03}	19
Propanoate metabolism	2.10×10^{-03}	12
Alanine, aspartate and glutamate metabolism	3.70×10^{-03}	12
Bile secretion	5.90×10^{-03}	17

C. Top over-expressed ($FC_{R-/R+} > 6$)	
Gene	$FC_{R-/R+}$
TENM2	49.94
CTNNA3	42.00
PCK1	19.93
SPATA4	9.74
GPC5	9.69
SULT1B1	8.27
LOC777017	8.15
LMBR1L	7.09
ATP6V0D2	6.61
APOA4	6.29

D. Top under-expressed ($FC_{R-/R+} < 1/6$)	
Gene	$FC_{R-/R+}$
CPNE4	0.015
RTN1	0.072
IGF2BP1	0.141

Figure 1: KEGG terms associated with the liver over- (A) and under- (B) expressed genes, and the top over- (C) or under- (D) expressed genes in R- versus R+, with TPM ≥ 1. KEGG terms are accompanied by the terms *p-values* and the number of associated genes; the top differentially expressed genes are accompanied by their fold-change.

In liver, we observed 3781 DE protein-coding genes with 1903 over-expressed and 1878 under-expressed genes in R- compared to R+. Enrichment analysis of the over- and under-expressed gene lists identified 11 and 15 significantly enriched pathways ($p_{FDR} < 0.05$), respectively (Additional file 1) and those with a $p_{FDR} < 0.01$ are indicated in Figure 1A and B. Pathways associated with the over-expressed genes were related to protein processing in the endoplasmic reticulum, Ribosome (23 ribosomal proteins L (RPL), 7 ribosomal proteins S and 4 mitochondrial ribosomal proteins: 3 L and 1 S) and protein export. Galactose metabolism also differed between the lines (*GALM*, *GALK1*, *GALT*, *GALE*), possibly in relation with the amino sugar and nucleotide sugar metabolism. In addition, numerous genes were associated with different metabolic pathways such as the phosphatidylinositol signaling pathway (*PI4K2A*, *PI4KA*, *PIK3C2B*, *MTMR4*, *MTMR2*), the glycolysis (*PFKL*, *PDHB*, *HK3*, *DLAT*, *LDHA*, *ADPGK*, *G6PC*, *HKDC1*, *GAPDH*), the fatty acids synthesis (*ME1*, *ACLY*, *ACACA*, *FASN*, *ELOVL6*, *FADS1*, *HSD17B12*), the transport and activation of fatty acids (*SLC27A4*, *FABP1*, *FABP7*, *ACSL1*, *ACSL3*, *ACSL4*), the triglycerides synthesis (*AGPAT5*, *AGPAT9* alias GPAT3), the lipid storage in droplets (*PLIN1, PLIN2, PLIN4, CIDEC, PNPLA3*) and lipid exportation (*MTTP*, *APOA1*, *APOA5*, *APOC3*). Ten genes had a fold-change in R- greater than 6 and a TPM $\geq$ 1 (Figure 1C). Among these, there were the *PCK1* involved in the gluconeogenesis and glyceroneogenesis and the *APOA4* involved in lipid transport.

The pathways associated with the under-expressed genes in R- compared to R+ were related to amino-acid catabolism such as tryptophan utilization. The glycine, serine and threonine metabolism was also affected, as indicated by the under-expression of *CHDH*, *ALDH7A1*, *DMGDH* and *PIPOX* genes all contributing to the formation of glycine from choline, of the *SHMT1* gene that acts to interconvert serine and glycine, and of the *AGXT* gene that converts the glyoxylate in glycine and the hydroxy-pyruvate in serine, which is the substrate of the enzyme coded by *SDSL* to form pyruvate. Many enzymes from the valine, leucine and isoleucine degradation pathways were also under-expressed. Within the alanine, aspartate and glutamate metabolism, *ADSL*, *ADSS*, *ASS1* are part of reactions that catalyze the formation of fumarate from aspartate, as an entry point into the citrate cycle. Under–expressed genes were also involved in cholesterol metabolism, more precisely in the cholesterol efflux, esterification and storage (*SCARB1*, *ACAT2*, *NCEH1*) and the transport (*ABCG5*, *ABCG8*) and synthesis (*CYP7A1*, *ABCB11*, *ABCC3*) of bile acid from cholesterol. The pyrimidine metabolism also seems to be impacted with *NME3*, *NT5M*, *NT5C3B*, *NT5C2*, *UPP2*, *DPYD*, and *CDA*, which leads to the synthesis of uridine. Three genes had a fold-change in R- lower than $1/6^{th}$ and a TPM $\geq$ 1 (Figure 1D). Among these gene is *IGF2BP1*, that was showed in *in vitro* analyses on

chicken liver-related cell line to regulate the expression of genes associated with fatty acid metabolism [32].

Functional characterization of adipose tissue transcriptome differences between the R+ and R- lines

Term	p_{FDR}	Nb Genes
A. Over-expressed in R- versus R+		
Oxidative phosphorylation	$3{,}59\times10^{-06}$	32
Metabolic pathways	4.40×10^{-06}	141
Thermogenesis	2.30×10^{-04}	38
Cardiac muscle contraction	1.10×10^{-03}	18
Proteasome	5.20×10^{-03}	12
Retrograde endocannabinoid signaling	8.20×10^{-03}	24
B. Under-expressed in R- versus R+		
None		

D. Top under-expressed in R- versus R+ ($FC_{R-/R+}$ < 1/6)

Gene	$FC_{R-/R+}$	Gene	$FC_{R-/R+}$
CCLI8	0.006	HTR1A	0.116
RAB3IP	0.010	TBX22	0.117
CES1L2	0.022	DLK1	0.132
ZDHHC22	0.048	TPPP3	0.134
GRXCR1	0.095	PCDH15	0.141
NETO2	0.097	DAW1	0.158
TMEM108	0.107		
COL8A2	0.116		

C. Top over-expressed in R- versus R+ ($FC_{R-/R+}$ < 1/6)

Gene	$FC_{R-/R+}$	Gene	$FC_{R-/R+}$	Gene	$FC_{R-/R+}$	Gene	$FC_{R-/R+}$	Gene	$FC_{R-/R+}$
ENSGALG{0×6}41713	58.7	ENSGALG{0×6}30700	13.9	PIT54	10.9	LOC107049102	7.8	LOC776376	6.2
AvBD5	38.5	CRP	13.0	AvBD10	9.4	SULT1B1	7.5	ENSGALG{0×6}11136	6.1
LOC770996	25.0	RSPO4	12.0	TDRD6	9.0	TUBA1C	7.3		
AvBD1	17.3	NTRK3	11.8	DPP10	8.6	APOV1	6.9		
GC	16.9	AvBD9	11.7	APOA4	8.5	LBFABP	6.9		
NPY	14.4	CL2	11.3	VTG2	8.4	ALB	6.4		

Figure 2: KEGG terms associated with the adipose tissue over- (A) and under- (B) expressed genes, and the top over- (C) or under- (D) expressed genes in R- versus R+. KEGG terms are accompanied by the terms *p-values* and the number of associated genes; the top differentially expressed genes are accompanied by their fold-change. For genes that had only an Ensembl ID (ENSGALG), the number of 0s in the ID is indicated in brackets to ease reading.

In the adipose tissue, we observed 3177 DE protein-coding genes with 1587 over-expressed and 1590 under-expressed genes in R- compared to R+. Enrichment analysis of the over- and under-expressed genes identified 15 and 1 significantly enriched pathways ($p_{FDR} < 0.05$), respectively (Additional file 1) and those with a $p_{FDR} < 0.01$ are indicated in Figure 2A and B, and the genes with the largest differential expression (fold-change absolute value ≥ 6 and TPM ≥ 1) are presented in Figure 2C and D. Among the pathways associated with the over-expressed genes in the R- line the oxidative phosphorylation, an energy metabolic process that produces ATP from ADP in the mitochondria, was the most significantly enriched. In this pathway the DE genes included 13 from the NADH:ubiquinone oxidoreductase, also known as mitochondrial complex I, (*NDUFS2*, *3* and *6*, *NDUFV2*, *NDUFA2*, *5*, *8* and *11*, *NDUFAB1*, *NDUFB2*, *5* and *9* and *NDUFC2*), 5 from the cytochrome C oxidase (*COX4I1*, *COX5A*, *COX6A1*, *COX6B1*, *COX11*), 3 from the cytochrome C reductase (*UQCRFS1*, *UQCRC1*, *UQCR11*), and 4 from the ATPase (*ATP6V1C1*, *ATP6V1G1*, *ATP6V0C*, *ATP6V0D1*). Concerning the enriched "metabolic pathways" KEGG terms, genes involved in the cholesterol

synthesis were increased (*CYP51A1*, *DHCR7*, *FDFT1*, *FDPS*, *GGPS1*, *LSS*, *PMVK*). Furthermore, genes related to the extracellular matrix (*COL1A1*, *COL4A1*, *COL4A2*, *COL4A3*, *COL9A3*, *LAMB1*, *LAMB4*, *THBS2*, *VTN, P4HA3*), genes from the integrin family (*ITGA3*, *ITGA5*, *ITGA7*), cytoskeleton related genes (*ACTG1*, *ACTN4*, *MYL12A*, *MYL12B*) and genes related to Gap junctions (Top 10 term, *TJP1*, *TUBA1C*, *TUBA1A*) were also over-expressed in the R- line. The APOA4 gene was among the most over-expressed genes in R-, as already observed in liver. Similarly, the NPY gene, that is an important central orexigenic hormone, was also among the top over-expressed genes, as several genes related to the immune-response including *CRP* which is an inflammatory protein, 3 avian β-defensins (*AvBD5*, *AvBD9*, *AvBD10*) that are antimicrobial peptides.

The only pathway associated with the under-expressed genes in R- compared to R+ was related to axon guidance. This pathway is supported by genes involved in cell adhesion and differentiation (NTNG2 and LRRC4C), in cytoskeleton cell motility and adhesion (*ABL1*, *L1CAM* and *EPHB2* and *6*) in cell-cell adhesion and cell guidance (*UNC5D*) , in microtubule assembly (*DPYSL2*) , and in cell migration in response to semaphorins (*PLXNB2*, *SEMA3D*). Among the most under-expressed genes in R-, *PCDH15* belongs to the protocadherin family. In addition, several genes having a role in the cytoskeleton or extra-cellular matrix were also under-expressed in the R- line (*DAW1*, *TPPP3*, *COL8A2*, and *GRXCR1*).

Analysis of liver's and adipose tissue's lipidomes

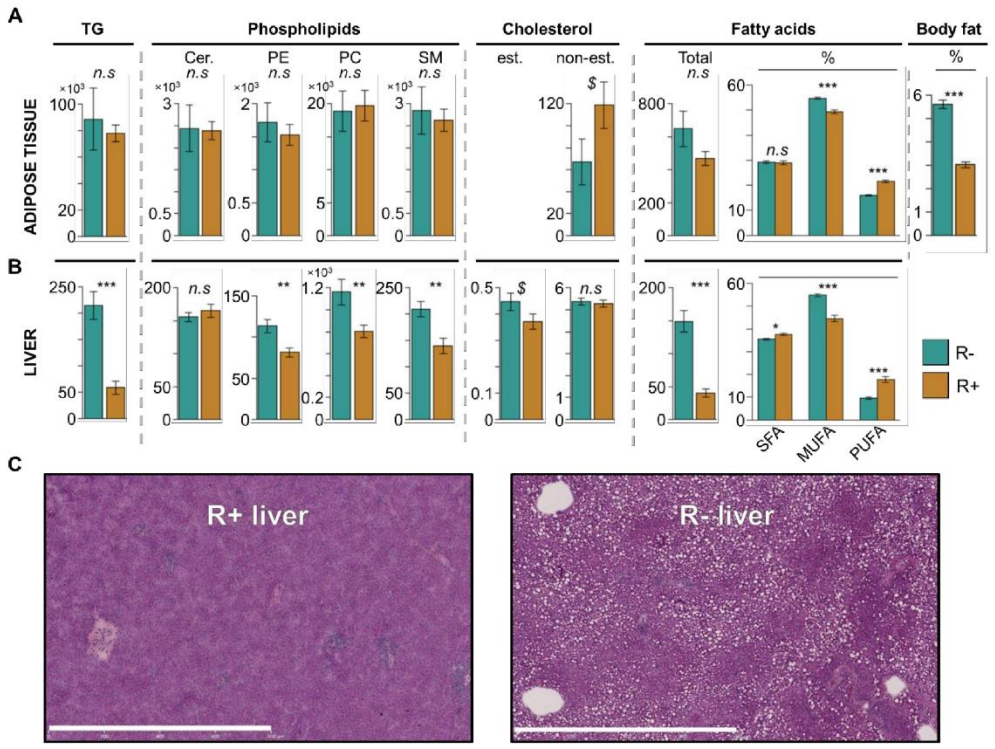


Figure 3: **Quantities of triglycerides, phospholipids, cholesterol and fatty acids in the adipose tissue (A) and liver (B), body fat indicator (top-right) and livers histological sections (C). A** and **B**.TG: triglycerides, Cer.: ceramides, PE: phosphor-ethanolamines, PC: phosphocholines, SM: sphingomyelins, est.: esterified, non-est.: non-esterified, all these lipids as well as the total fatty acids are expressed in relative abundance per mg of proteins (see Material and Methods). The "%" in fatty acids corresponds to the percentages of saturated, mono-unsaturated and poly-unsaturated fatty acids (SFA, MUFA and PUFA, respectively). %: percentages, SFA: saturated fatty acids, MUFA: mono-unsaturated fatty acids, PUFA: poly-unsaturated fatty acids. The top-right plot shows the percentage of abdominal adipose tissue mass to total body weight. Error-bars represent the standard error of the mean. *n.s*: not significant, $: *p-value* ≤ 0.1, *: *p-value* ≤ 0.05, **: *p-value* ≤ 0.01, ***: *p-value* ≤ 0.001. **C** Haematoxylin and Eosin staining of liver sections. Scale bars, 800 µm.

In complement to the transcriptome analysis, we analyzed the liver and adipose tissue lipidomes of both lines (Additional file 2). While the R- line had a larger proportion of abdominal fat on body weight compared to the R+ (right plot of Figure 3A), the lipid profile analysis from both tissues revealed almost no qualitative differences in fat composition between lines, while quantitative differences existed but mostly in liver (Figure 3B). Higher values in R- compared to R+ were observed for most of the triglycerides, phospholipids, cholesterol and fatty acids

analyzed (see Additional file 2 and Figure 3B). The liver fatty composition of the R- was compatible with a hepatic steatosis condition, with approximately 3 times more fatty acids per mg of liver and 3.6 times more triglycerides than the R+. We also observed 1.25 to 1.5 times more phospholipids and no difference in ceramides. The fatty acid composition of the R- included a higher percentage of mono-unsaturated fatty acids (MUFA) and a lower percentage of poly-unsaturated fatty acids (PUFA) in both tissues, and the ratio PUFA / MUFA was twice higher in R-. In addition, in the adipose tissue, we observed a tendency to a higher quantity of non-esterified cholesterol in the R+ line (119 in R+ versus 67 in R-, $p \leq 0.10$). In liver we observed a tendency to a lower quantity of esterified cholesterol in the R+ (0.37 in R+ versus 0.45 in R-, $p \leq 0.10$), while the quantity of non-esterified cholesterol remained unchanged. These differences were clearly visible in liver sections (Figure 3C), with the R- liver showing much more lipids.

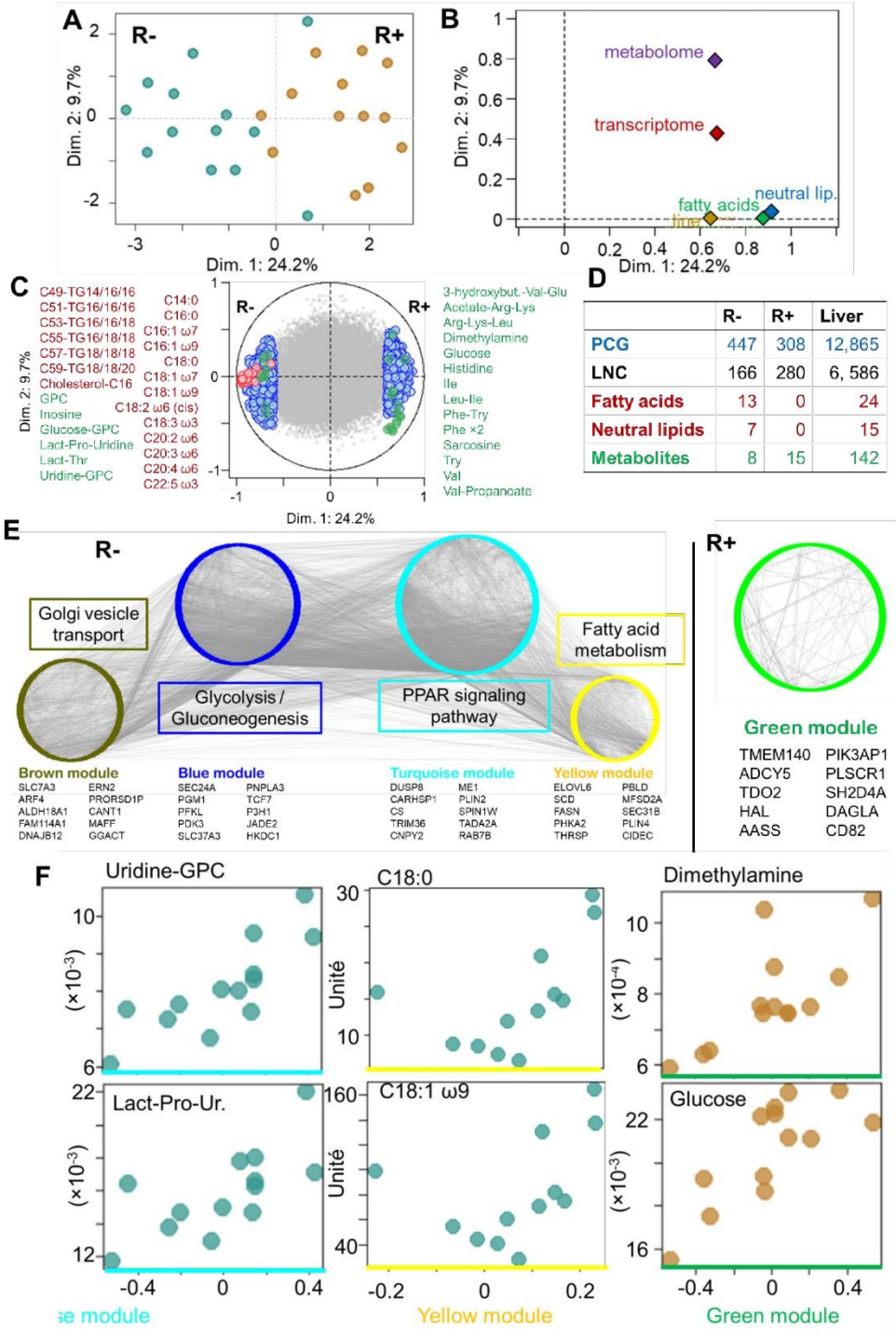
Integrative analysis of the transcriptomic, lipidomic and metabolomic data in the liver

Since the previous analysis showed that many DE genes in the liver were related to the metabolic pathways (152 over-expressed genes, and 195 under-expressed genes), we conjointly studied three types of data collected in the liver of 24 samples (12 per line): the transcriptome (19 612 variables in total), the lipidome (15 to 24 variables) and the metabolome (142 variables). Regarding the metabolome, on the 142 buckets detected, we found that 54 were significantly different between lines (for example the bile acids, 1.32× higher in R- versus R+, $p_{FDR} = 0.04$), 9 of which were associated to unknown molecules (Additional file 3). Interestingly, most of the significant buckets (39 out of 54, i.e. 72%) had a higher value in R+ versus R-.

The joint study aimed (*i*) at finding the variables that separate the best the lines and (*ii*) at relating the variables of these three datasets, in particular to identify the genes most related to some lipidomic or metabolomic variables.

To this end, we realized a Multiple Factors Analysis (MFA), a method well adapted for simultaneous analysis of multiple datasets. It ensures that all datasets play an equal role in the analysis regardless of the number of variables they contain. We then studied the genes that contributed the most to the separation of the lines using Weighted correlation network analysis (WGCNA) R package. This method identifies groups of highly correlated genes (“modules”) and allows studying the correlation of the lipidomic or metabolomic variables to these modules, but also identifying the genes that that have the highest number of correlated genes, within and outside each module.

The results of the MFA using liver transcriptomic, lipidomic and metabolomic data are displayed in Figure 4.



	R-	R+	Liver
PCG	447	308	12,865
LNC	166	280	6, 586
Fatty acids	13	0	24
Neutral lipids	7	0	15
Metabolites	8	15	142

Figure 4: **Liver's transcriptome, lipidome and metabolome joint analysis by MFA and WGCNA gene clustering.** (**A**) Graph of the individuals from the MFA. The dimension 1 clearly separates the two lines. (**B**) Correlations of each datasets with the dimensions 1 and 2. All the datasets have a correlation ≥ 0.6 with dimension 1. (**C**) Variables responsible for the separation of the samples on axis 1 ($p \leq 0.001$). On the left (right) side the variables that are highly expressed in the R- (R+) are displayed. Blue points represent genes, red points represent lipids and green points represent metabolites. (**D**) Numbers genes, fatty acids, neutral lipids and metabolites responsible for the separation of the samples on axis 1 and positively correlated with R- (1st column), R+ (2nd column). The total number of analyzed variables per dataset is given in the 3rd column. (**E**) Gene modules detected by WGCNA in R- (left) and R+ (right). The name of each module corresponds to the most significant KEGG term enriched in the module's gene (when such enrichment exists). For each module, the top 10 most connected genes with other genes of the module are indicated. Grey lines represent the connection between all the genes. (**F**) Examples of correlation between the indicated module's eigengene (x-axis) and some of the lipids or metabolites (y-axis). GPC: glycerophosphocholine.

The MFA 1st dimension provided a good separation between the R- and R+ lines as shown in Figure 4A. Every dataset participated to this separation, as shown Figure 4B which represents the correlation of each datasets to dimension 1 and 2. All datasets had a correlation greater than 0.6 with dimension 1. On the 755 genes that were significantly correlated to this dimension ($p \leq 0.001$, $| r | > 0.6$), 308 were over-expressed in the R+ and 447 were over-expressed in the R- (Figure 4C and D). As expected, all lipidomic variables that correlated with dimension 1 were more abundant in R- (Figure 4C). In contrast, of the 13 metabolites that were correlated with dimension 1, 8 were more abundant in R- and 15 in R+.

To identify cluster of highly correlated genes (thereafter called "modules") within the two gene lists identified by MFA and to identify the metabolomics and lipidomic variables the most correlated to these modules, we performed a WGCNA analysis (Figure 4E). In total we identified 4 modules (Brown, Blue, Turquoise and Yellow) comprising 85 to 231 genes among the genes highly expressed in R-, and 1 module (Green module) including 483 genes, among the genes highly expressed in R+. As expected, the four R- modules were enriched in genes associated to different pathways already identified in the pathway enrichment analysis of DE genes, and the top of which (based on *p-values*) were annotated by each module Figure 4E. The vesicular transport in the Golgi apparatus enriched term of the Brown module was supported by *ARCN1*, *ARF4*, *COG1*, *COPB2*, *COPE*, *KDELR2*, *RAB6A*, *RER1*, *SEC13*, *SEC31A* and *USO1*, the Blue module Glycolysis & Gluconeogenesis enriched pathway by *BPGM*, *HKDC1*, *PDHB*, *PFKL* and *PGM1*, the Turquoise module PPAR signaling pathway by *ACSL4*, *ME1*, *PCK1*, *PLIN1* and *PLIN2*, and the Yellow module Fatty acid metabolism was supported by

ELOVL6, *FASN*, *HSD17B12* and *SCD*. On the other hand, no enrichment was detected in the R+ module. In order to search for potential regulators, we screened in each module for the top 10 genes with the highest number of connections with other genes of the same module. These groups of 10 genes are indicated in Figure 4E.

Finally, in order to relate the genes modules with the other analyzed components, we studied the correlations between each module eigengene, which is a weighted average of expression of all genes in a given module, with the lipids and metabolites analyzed. Figure 4F shows the most significant correlations for the modules identified for the R- ($p < 0.01$ in R-) and R+ ($p < 0.05$ in R+). The R- the Turquoise module's eigengene was significantly correlated with glycerophosphocholine (GPC) and Lactate-Proline-Uridine. The Yellow module eigengene was significantly correlated with the SFA C18:0 and its product C18:1 ω9 obtained after desaturation with *SCD* (top 2 most connected gene). In the R+, the Green module eigengene was significantly correlated with Dimethylamine and Glucose.

Discussion

This study considered layers from two genetic chicken lines divergently selected for RFI, a statistical-built estimate related to the efficiency of feed utilization. Previous studies have advanced the hypothesis that the selection for a low or a high RFI determined a change on important metabolic functions [4-7,41]. Our study allows to explore the functional impact of selection and to propose hypothesis to explain the observed differences.

R- line characterized by *de novo* lipogenesis, hepatic steatosis and high adiposity

In poultry diets, carbohydrates (mainly starch) are the main source of energy, supplying about 50% of the energy while the dietary lipid energy is often less than 10 % [33]. In birds, body lipids are mainly synthesized *de novo* from carbohydrates in the liver [8]. Our results indicate an enhancement of this *de novo* fatty acid synthesis in the R- compared to the R+.

The transcriptional pattern observed between lines strongly suggests an increased hepatic glycolysis in the R- line since numerous genes involved in the metabolic pathway that convert glucose into pyruvate were over-expressed in this line. Among these, there were the *HKDC1* gene that codes for a hexokinase isoform involved in the first step of glycolysis [34] and the *PDHB* and *DLAT* genes that code for subunits of the pyruvate dehydrogenase complex, which produces acetyl-CoA that is the molecule linking glycolysis and *de novo* lipogenesis through the Krebs cycle [35] as discussed afterwards.

We then observed in the liver of the R- line an over-expression of genes encoding key enzymes intervening at different levels of lipid metabolism as the fatty acid biosynthesis process, the transport and activation of fatty acids, the triglyceride synthesis, the lipid storage in droplets and the lipid exportation to peripheral tissues through lipoproteins. The WGCNA analysis of the genes found by the MFA that mostly contributed in the separation of the two lines highlighted the Yellow module, which was enriched in genes related to fatty acid synthesis. In particular, the three most connected genes of this module encodes some of the most important enzymes of this process: *FASN*, which produces palmitate (C16:0), from malonyl-CoA, *ELOVL6*, which elongates palmitoyl-CoA to stearoyl-CoA (C18:0) and SCD, which desaturates saturated fatty acids (SFA) to mono-unsaturated fatty acids (MUFA). In previous studies the expression of *SCD* was positively associated with adiposity [36–38]. We observed a similar positive association in the R- that presented larger abdominal fat content and, at the liver level, higher amounts of total fatty acid and MUFA compared to the R+. Interestingly, studies on mice showed that deficient *SCD* mice (*SCD*-/-) had a reduced hepatic triglycerides and cholesterol esters [39] despite an increased activity of the *ACAT* enzyme that catalases the synthesis of cholesterol esters, which is in agreement with the expression profile and lipidome results that we observed in the R+ line. Moreover, in the *SCD*-/- mice increased energy expenditure, reduced body adiposity and increased insulin sensitivity were also observed [40,41], which are all traits that characterize the R+ line [4,7,42]. It is highly probable, in the light of these similarities between the mice *SCD* -/- model and the R+ line that *SCD* gene is an important metabolic control point in the divergence of lipid related traits between lines. Previous studies on these lines showed that when fed deprived, R+ birds had a higher plasma concentration of non-esterified FA than the R- and this could suggest that FA in the R+ are used as fuel to enhance thermogenesis [5]. The same study showed also that, in the fed state, the R- chickens presented low plasma T3 associated to high plasma insulin, an endocrine environment that would reduce thermogenesis expenditure and stimulate lipogenesis, respectively.

In the R- the liver metabolic machinery is programed to synthesize *de novo* lipids. The fate of these lipids is, at first, to be stored as lipid droplets (LD) in the cytosol of the hepatocytes. This is suggested by the over-expression of several lipid droplet-specific genes in the R- line(*PLIN1*, *PLIN2*, *PLIN4*, *CIDEC* and *PNPLA3*) that are necessary to maintain LD homeostasis [43] and are markers of steatosis [44–46] and also of numerous genes related to “protein processing in endoplasmic reticulum” (ER) (Figure 1A), the ER being the place of formation of the LD [43]. In a second time, these lipids are most likely exported, to other organs, largely to the adipose

tissue, as suggested by the over-expression of *MTTP* which plays an important role in lipoproteins assembly [47], and of genes coding for apolipoproteins, such as *APOA1* [48], *APOA4* [49], *APOA5* [50] and *APOC3* [51] (all DE with $p_{FDR} \leq 10^{-3}$), although the precise roles of each of these proteins remain to be fully elucidated [52,53].

The increased *de novo* lipogenesis in liver and exportation to peripheral tissue is consistent with the accumulation of lipids in the liver (steatosis) and consequent ectopic lipid accumulation that could signal an adipose tissue dysfunction related to a lipid overflow [54].

A complex gene regulation behind the de novo lipogenesis in the R- liver

Regarding the regulation of *de novo* lipogenesis in the liver, we observed that *NR1H3* (alias *LXRα*), which codes for a receptor involved in the control of various physiological functions with a major role in fatty acid homeostasis and cholesterol metabolism [55] was not differentially expressed.

The over-expression of *de novo* lipogenesis genes as *SCD*, *FASN*, *ACACA* and *ELOVL6*, which are target genes of SREBP1, a key transcription factor in the lipogenesis in different species of mammals [56] and in chicken [57,58], made us assume that *SREBF1* was also differentially expressed. Indeed, in previous studies conducted in chicken in which these different lipogenic genes were transcriptionally regulated in liver, the expression of *SREBF1* was systematically impacted [8,59]. However, this was not the case in this present study. Instead we found that *SCAP* (Sterol regulatory element-binding protein cleavage-activating protein) [60], a functionally related gene that cleaves proteolytically inactive SREBP1 (*SREBF1* encoded protein) making it transcriptionally active [61,62], was over-expressed in R- suggesting an activation of SREBP1 by posttranscriptional mechanism.

Lipogenesis is tightly regulated at the transcriptional level. *THRSP (alias SPOT14)* also appears to be necessary for *de novo* fatty acid synthesis and was over-expressed in the R- line liver, and was among the top 10 most connected genes of the yellow module that was enriched in genes related to fatty acid synthesis. Even if the precise mechanism remains unclear, this gene is highly expressed in tissues with high rates of lipogenesis [63], acts in the induction of mRNAs coding for key lipogenic enzyme [64], and the knockout mice have reduced fatty acid synthesis in lactating mammary glands [55,56].

Another important gene involved in metabolism and energy homeostasis is *FOXO1* that was over-expressed in R-. It is an important downstream mediator of the insulin signaling pathway that, in a context of insulin resistance, contributes to the accumulation of hepatic lipids and to an important production of apolipoprotein C3 [65–67], which coding gene was over-expressed

in R- ($FC_{R-/R+}$ = 1.87, $p_{FDR} = 1.80\times10^{-5}$). According to Dong *et al.*, 2008 [68], *FOXO1* regulates the expression of different genes that we also find to be differentially expressed in our study, among which there are *PCK1* and *G6PC* involved in the gluconeogenesis both over-expressed in the R- ($FC_{R-/R+}$ = 19.93, $FC_{R-/R+}$ = 1.30, respectively) and *MTTP*, encoding the microsomal triglyceride transfer protein that facilitates the transport of dietary and endogenous fat, also over-expressed in the R- as already mentioned above. *G6PC* codes for a key enzyme that catalyzes the final step in gluconeogenesis, the conversion of glucose 6-phosphate to glucose [69]. *PCK1* codes for the phosphoenolpyruvate carboxykinase that catalyzes the first rate-limiting step in hepatic gluconeogenesis pathway to maintain blood glucose levels. This transcriptional profile indicate that liver glucose synthesis is higher in R- than in R+, which, together with the previously described lipogenesis pattern, could suggest that selection on RFI is associated with changes in the insulin sensitivity leading to a liver insulin resistant condition for the R-. Moreover, Kamagate *et al*. (2008) showed that overexpression of *FOXO1* also increases *MTTP* expression and this increased expression was inhibited by insulin [70]. Variation in glucose-insulin relationships were also identified in previous studies on the R+ and R- lines showing that plasma glucose concentrations were higher in the R- line compared with the R+ when fasted, whereas R- birds exhibited a higher plasma insulin concentration than R+ birds either in fed or fasted state [42]. Moreover, insulin resistant condition is often associated to NAFLD, which is common in R- animals. This causes insulin-resistant individuals to fail to inhibit hepatic glucose production and have increased liver lipid synthesis [71] leading to liver dysfunction [72]. We can note that *PCK1* could have a second role in glyceroneogenesis, that occurs in the liver along with gluconeogenesis and provides glycerol-3-phosphate (G3P), necessary for triglycerides synthesis, from non-carbohydrate precursors (lactate, pyruvate and amino acids) for lipogenesis [73,74].

Alterations in cholesterol and/or bile acid metabolism were observed between lines in liver and adipose tissue

In liver, the under-expressed genes in R- compared to R+ were enriched in genes associated with bile acids synthesis and excretion. Among these, there was CYP7A1 that uses as substrate cholesterol from cholesterol esters [75] and is the first and the rate-limiting enzyme in the classic bile acid biosynthesis in liver [76,77]. The genes coding for the subsequent enzymes of this pathways were also differentially expressed, or tended to: *HSD3B7*, *AKR1D1*, *CYP8B1* [78]. Furthermore, genes involved in the excretion of the bile acids bile into the canaliculus (*ABCB11*) and / or the circulation (*ABCC3*) were also under-expressed. These observations at

the transcriptomic level are consistent, at the lipidomic level, with the higher concentration of esterified cholesterol in the R- liver (0.37 in R+ versus 0.45 in R-, $p \leq 0.10$), since esterified cholesterol is the substrate at the beginning of the synthesis pathway. The under-expression of the genes coding for the bile-exporting proteins may explain the observation at the metabolomic level of a higher concentration of bile acids in the R- liver. In turn, this impaired exportation may be responsible for the under-expression of CYP7A1, since bile acids negatively regulate their own biosynthesis [79] or to the action of insulin, which suppresses bile acid synthesis by down-regulating CYP7A1 [80]. At the organism level, the increased hepatic bile acid production and exportation of the R+ line may be related to the hyperphagia of this line as both hepatic bile acid synthesis and plasma bile acid concentrations increase in response to food intake [81]. In addition, higher expression of hepatic *CYP7A1* and bile acid production in R+ may help prevent diet-induced obesity, fatty liver and insulin resistance, as previously reported in mice [82]. Dysregulation of bile acids synthesis and exportation are also consistent with observations in human affected by non-alcoholic fatty liver disease (NAFLD) [83,84].

In adipose tissue, many genes involved in cholesterol synthesis were over-expressed in the R- line characterized by a higher abdominal adipose tissue mass compared to the R+. Such results are consistent with previous studies in mice or human models since cholesterol accumulates in large quantities in a non-esterified form in adipose tissue particularly within the adipocyte lipid droplet, that major organelle allowing an efficient energy storage in adipose tissue (for review see [85,86]). The expansion of lipid droplets in adipocyte requires a redistribution of the cholesterol from the plasma membrane to lipid droplet membranes. This cholesterol deprivation of plasma membrane in the adipocytes induces cholesterol synthesis, through activation of different genes targets of SREBP2, the major transcription factor of the cholesterol synthesis [87]. In our model, *SREBF2* is not differential expressed between lines but, as for SREBP1 that regulates the lipogenesis in the liver, the activation of the SREBP2 in adipose tissue could occur through the SREBP cleavage by the SCAP protein, the associated gene being significantly overexpressed in the adipose tissue of the R- line.

The R- adipose tissue shows signs of cholesterol synthesis, mitochondrial dysfunction, inflammation and remodeling

The over-expressed genes in R- adipose tissue were enriched in terms linked to the oxidative phosphorylation, and particularly of genes coding for proteins composing the complex I of the mitochondrial respiratory chain (NADH:ubiquinone oxidoreductase), which is a major

producer of Reactive Oxygen Species (ROS) [88]. Consistently with an increase in ROS production, we observed tendency for an over-expression of the genes encoding the antioxidant enzyme *SOD1* ($FC_{R-/R+} = 1.10$, $p_{FDR} = 0.08$). Interestingly, we observed a tendency to the under-expression of the gene encoding the antioxidant enzyme *GSTA4* ($FC_{R-/R+} = 0.45$, $p_{FDR} = 0.06$), as was observed in obese insulin-resistant mice and in human [89].

We further observed an over-expression of *CRP*, an inflammatory protein that is a circulating marker of low-grade systemic inflammation [90] that can be synthesized in the adipose tissue under pro-inflammatory conditions [91]. In human, these conditions usually imply the expression of *IL1-β* or *IL6*. In our data, *IL1-β* and *HSF3* that could play the role of interleukin 6 which is absent in chicken [92], and *APP* another inflammation marker tends to be significant over-expressed in the R- line ($p_{FDR} < 0.1$). Interestingly we observed in the R- line an over-expression of all the 5 avian β-defensin genes present in our dataset. Defensins have important antimicrobial activity but immunomodulatory properties, including anti-inflammatory properties, have also been demonstrated [93–95], and in vivo, up-regulation of β-defensin genes has been shown to occur in both infectious and inflammatory states [96]. β-defensin is the only type of defensin found in avian species [97], and 14 of them are currently identified in chicken [97]. The over-expression observed in the R- lines may reflect an increased inflammation in the adipose tissue of this line. This would be due to the fact that an increase in lipid uptake in the adipocyte induces an increase in mitochondrial substrate load, leading to an overloading, inducing ROS production and inflammation [98].

Many genes composing the adipocyte extracellular matrix ECM [99] were differentially-expressed in the R- adipose tissue, as was observed in previous studies in obese human adipocytes [100]. Among these over-expressed genes there were eleven genes from the collagen family and three transmembrane proteoglycans that regulate inflammatory responses [101]. These changes may be linked to a remodeling of the adipose tissue in the R- line driven by the inflammation or by a constant lipid uptake or both. It is in fact known that the extracellular matrix (ECM) is central for adipose tissue remodeling and in response to the inflammation, the adipose tissue undergoes functional and morphological changes, in order to accommodate with the input of lipids to store [102]. In addition, under lipid uptake the adipocytes size or number increase inducing ECM adjustments [103]. Finally, it appears that insulin also promotes the expression of enzymes involved in the formation components of the ECM [99]. Among the over-expressed ECM related genes there was *P4HA3*, that codes for a component of prolyl 4-hydroxylase [104], responsible for the modification of collagens necessary for their assembly into collagen monomers and fibrils, that also seems to be positively regulated by insulin [105].

The regulations behind this remodeling are likely to be diverse. Among the most differentially expressed genes in the adipose tissue there were *PRLL* (prolactin-like protein) and *NPY* (neuropeptide Y), both having an important role in adipose tissue accumulation. The prolactin gene, *PRL*, and its receptor, *PRLR*, act on different aspects of the adipose tissue metabolism. For examples, females mice with a *PRLR* knock-out have reduced adipose tissue weight [106], through reduced number of adipocytes (but not of their volume [107]), and in human, *PRLR* was shown to inhibit lipoprotein lipase activity [108]. Chicken *PRLL* shares 30 to 35% amino acid sequence identity with chicken prolactin (*PRL*) [109] and has been shown to activate chicken *PRLR* [110]. The second highly expressed gene, *NPY*, was shown to be synthetized in adipose tissue, where it promotes proliferation of adipocyte precursor cells [111] and therefore lipid accumulation [112]. The same effects on adipogenesis were observed in vitro in chicken adipocytes [113]. Interestingly, Kuo *et al*. [114] showed that under chronic stress, both *NPY* and *NPY2R* (a peripheral receptor of *NPY*) are up-regulated in adipocytes. This creates a positive feedback that increases the proliferation and differentiation of new adipocytes, hence the growth of abdominal fat [114]. In addition, conditional knock-down of *NPY2R* lead to a reduction in adiposity and weight gain in mice when fed a high-fat diet, and a reduction in energy expenditure when fed a chow diet [115]. These results suggest that in the R- line both *PRLL* and *NPY* may be involved in the development and maintenance of an important adipose tissue in the R- line.

Finally, it is likely that a crosstalk exists between the adipose tissue and the liver. This crosstalk may be in part mediated in the R- by the adiponectin (*ADIPOQ*) a secretory protein synthetized by the adipose tissue, which exerts effect in different tissues. Adiponectin increases metabolic flexibility of adipose tissue, hence enhancing its ability to maintain proper function under metabolically challenging conditions [116], and decreased levels of this hormone are markers of metabolic disorders such as the metabolic syndrome [117]. The liver is a target organ for adiponectin, and we observed an over-expression of the adiponectin receptor 2 (ADIPOR2,) in the liver of R- line. Importantly, a large share of the adiponectin synthesized is not secreted, but retained in the adipocytes. The control of the release of this hormone from the adipocytes is posttranslational [118].

Conclusions

Organism-level regulation of the energy homeostasis involves numerous, complex, regulatory mechanisms in many tissues, that are not entirely deciphered. In addition, these mechanisms act at different levels (expression level, post-transcriptional and -translational levels, etc.) making it difficult to study them all at once. In chicken, the energy metabolism has long been a selection target to improve efficiency but the underlying genetic basis regulating it are still extremely scarce. In this work, we exposed the effects on the adipose tissue and liver transcriptomes and lipidomes of more than 40 years of divergent selection on feed-efficiency. We show that the transcriptional and metabolic profiles of the R- line are compatible with an insulin-resistance condition, characterized by an important hepatic *de novo* lipogenesis and steatosis, and an accumulation of lipid in the adipose tissue, probably undergoing inflammation and remodeling processes. We also proposed some mechanisms that could be involved in the regulations of the observed phenomenon, with a focus on *SCAP* and its cleavage of SREBP1, *THSRP*, *FOXO1* and *ADIPOQ*. Studying evolutionary conserved mechanisms involved in these diseases may be of help for a better understanding of these conditions in human.

Material and methods

Animals

Laying hens were hatched at the INRAE *Pôle d'Expérimentation Avicole de Tours* (PEAT) in Nouzilly, France. They belonged to two Rhode Island Red layer lines that underwent a 40-year diverging selection on residual feed intake (RFI) [3]. The RFI represents the difference between the observed and the predicted feed consumption based on a multiple regression equation taking into account the average body weight, the weight variation and, for females, the mass of eggs produced over a given period [1,119]. The R+ chickens were selected to have a positive RFI, reflecting a low feed efficiency, while the R- chickens were selected to have a negative RFI and therefore to be feed efficient. For the transcriptomic data, we used liver samples from 24 R+ and 24 R- from 3 conditions (n=8): control (CT), low-energy diet (LE, also used in [120]) and heat-stress (HS). For the adipose tissue, we used samples from the same birds as previously, but only from CT and LE conditions for which the tissue was collected. No interaction between the line and the conditions was found in any tissue, leading us to gather all the birds to increase the statistical power for the line differential analysis. Regarding the lipidome, we used for the adipose tissue, birds from the two lines in the CT group (24 birds, n=12) and for the liver, birds from the two lines in the CT and HS groups (48 birds, n=12). The metabolomic data were produced on 72 birds from the two lines and the three conditions CT, HS and LE (n=12). The

integrative analysis in liver was undertaken on 26 birds of the CT and HS groups (13 from each line) for which all the transcriptomic, lipidomic and metabolomic data were available.

Tissue sampling

At 31 weeks, 48 hens (8 animals per line and conditions: control CT, low-energy diet LE, heat-stress HS) were selected as representative of the experimental population. Hens were slaughtered at the fed status by neck cut and bleeding, immediately after electrical stunning. Right after slaughter, abdominal adipose tissue from the CT and LE groups and the extremity of the left liver lobe from the CT, LE and HS groups were sampled, snap frozen in liquid nitrogen and stored at -80 °C until analysis.

RNA isolation

RNA extraction was performed on 32 adipose tissue samples and 48 liver samples. Approximately 100 mg of adipose tissue and 30 mg of liver were homogenized in TRIzol® reagent (Invitrogen, California, USA). The total RNA was then extracted according to the manufacturer's instructions, resuspended in 50 µL of RNA-free water and stored at − 80 °C. The total RNA was quantified with a NanoDrop® ND-1000 spectrophotometer (Thermo Scientific, Illkirch, France). The RNA quality was controlled using an Agilent 2100 bioanalyzer (Agilent Technologies France, Massy, France). The average RNA integrity numbers were 7.3 ± 0.6 (mean ± SD) for the adipose tissue and 9.2 ± 0.3 for the liver.

RNA-seq data acquisition

Paired-end sequencing was conducted on all samples using an Illumina HiSeq3000 (Illumina, California, USA) system, with 2×150 bp. Libraries with an average 508-bp insert were prepared following Illumina's instructions by purifying poly-A RNAs (TruSeq RNA Sample Prep Kit). Illumina adapters containing indexing tags were added for subsequent identification of samples. Samples were PCR-amplified, and quantitative PCR was then performed for library quantification (QPCR NGS Library Quantification kit). Eight samples were filled on one lane within a flow cell with 2 samples for each of the four line × diet or line × temperature groups to minimize the inter-lane bias. After sequencing, the indexed adapter sequences were trimmed using CASAVA v.1.8.2 software (Illumina). We obtained an average of 90 million reads per sample (85 million for the adipose tissue and 86 million for the liver), for a grand total of 7 billion reads. For each sample, reads were mapped to the Gallus gallus-5 reference genome using STAR v.2.3.0e [121]. PCR duplicates were removed using rmdup tool from SAMtools

suite[122]. For each sample, quantification was performed using RSEM [123] with the Ensembl v93 annotation.

RNA-seq data analysis

All the analyses were performed with R version 3.4.2 [124]. In each tissue, the expressed genes were selected if their TPM expressions were over 0.1 in at least 80% of the samples in one of the six modalities (R+, R-) × (CT, HS or LE). Differential expression analysis was performed using the R/Bioconductor package edgeR [125] version 3.16.5, based on a generalized negative binomial model for model fitting taking into account the two factors (line × condition) and interaction between the two factor (found as non-significant). We used the "edgeR-Robust" method to account for potential outliers when estimating per gene dispersion parameters[126]. Expression are presented as RPKM, outputted from the "rpkm" function of the edgeR package, in which the trimmed mean of M-values (TMM) scaling factor method is used for library size normalization [127]. P-values were corrected for multiple testing using the Benjamini-Hochberg approach [128] to control the false discovery rate (FDR), and genes were identified as significantly differentially expressed if $p_{FDR} < 0.05$.

Functional enrichment analysis

The enrichment analysis of Kyoto Encyclopedia of Genes and Genomes (KEGG) terms in each list of interest of differentially expressed genes was performed using the STRING tool [129] (https://string-db.org). Only the 1-to-1 human orthologous genes with a standardized HGNC name were submitted for the analysis, i.e. 67.4% of the 18,346 protein-coding genes of chicken Ensembl v93 annotation.

Analysis of liver neutral lipids

Lipids were analysed as previously described20. Tissue samples were homogenized in methanol/5 mM EGTA (2:1, v/v), and lipids (corresponding to an equivalent of 2 mg tissue) extracted according to the Bligh–Dyer method63, with chloroform/methanol/water (2.5:2.5:2 v/v/v), in the presence of the following internal standards: glyceryl trinonadecanoate, stigmasterol, and cholesteryl heptadecanoate (Sigma). Triglycerides, free cholesterol, and cholesterol esters were analysed by gas-liquid chromatography on a Focus Thermo Electron system equipped with a Zebron- 1 Phenomenex fused-silica capillary column (5 m, 0.25 mm i.d., 0.25 mm film thickness). The oven temperature was programmed to increase from 200 to

350 °C at 5 °C/min, and the carrier gas was hydrogen (0.5 bar). The injector and detector temperatures were 315 °C and 345 °C, respectively.

Liver fatty acid analysis

To measure all hepatic fatty acid methyl ester (FAME) molecular species, lipids that corresponded to an equivalent of 1 mg of liver were extracted in the presence of the internal standard, glyceryl triheptadecanoate (2 µg). The lipid extract was transmethylated with 1 ml BF3 in methanol (14% solution; Sigma) and 1 ml heptane for 60 min at 80 °C, and evaporated to dryness. The FAMEs were extracted with heptane/water (2:1). The organic phase was evaporated to dryness and dissolved in 50 µl ethyl acetate. A sample (1 µl) of total FAME was analysed by gas-liquid chromatography (Clarus 600 Perkin Elmer system, with Famewax RESTEK fused silica capillary columns, 30-m × 0.32-mm i.d., 0.25-µm film thickness). The oven temperature was programmed to increase from 110 °C to 220 °C at a rate of 2 °C/min, and the carrier gas was hydrogen (7.25 psi). The injector and detector temperatures were 225 °C and 245 °C, respectively.

Metabolite extraction and ^{1}H NMR analysis

NMR spectroscopy was performed on aqueous liver extracts prepared from liver samples (95–125 mg) in dichloromethane/methanol/water (2:2:2 v/v/v) as previously described in (REF). The ^{1}H NMR metabolomics analysis was performed by the AXIOM metabolomics platform (MetaToul) and spectra were obtained on a Bruker Avance III HD NMR spectrometer (Bruker) operating at 600.13 MHz for ^{1}H resonance frequency and equipped of a cryoprobe attached to a cryoplatform (the preamplifier cooling unit). The ^{1}H-NMR spectra were acquired at 300 K using a standard one-dimensional noesypr1D pulse sequence with water presaturation and a mixing time of 100 ms. A total of 256 free induction decays were collected into 32.768 data points using a spectral width of 20 ppm, a relaxation delay of 2 seconds and an acquisition time of 8 minutes. All ^{1}H spectra were subjected to 0.3 Hz exponential line broadening before Fourier transformation. The spectra were automatically phase and baseline corrected using Bruker Topspin 3.2 software (Bruker GmbH, Karlsruhe, Germany). The ^{1}H NMR spectra were normalized to the sum of total spectrum intensity to minimize the effect of the differences in sample concentration.

Lipids and metabolites statistical analysis

Analysis were carried out with R version 3.4.2. A two-way analysis of variance was performed with line, condition (CT, LE or HS) and the interaction between line and condition using the R function lm, and the R package "car" [130].

Multiple factor analysis (MFA)

We realized a MFA [131] using R package FactoMineR [132] to get a simultaneous view of transcriptomic, lipidomic and metabolomic differences between the lines and to detect the variables that best separate the two lines. MFA is an extension of principal component analysis (PCA) tailored to handle multiple data tables that measure sets of variables collected on the same observations [131]. Briefly, the MFA consists in the three following steps: 1) it computes a principal component analysis (PCA) of each data table and weights each table by dividing all table elements by largest eigenvalue obtained, in order to balance the contributions of each dataset, whatever its size (number of variables). 2) All the weighted data tables are aggregated into a grand data table that is analyzed via a non-normalized PCA that gives a set of factor scores for the observations and loadings, or correlations for the variables. 3) In addition, MFA provides for each data table a set of partial factor scores for the observations that reflects the specific 'view-point' of this data table. A threshold $r > |0.75|$ ($p\text{-}value < 0.001$) was used to extract these relevant variables as we did previously [8]. To facilitate the interpretation of the MFA results, the line factor was used in the MFA as illustrative variables (it does not contribute to the MFA components construction).

Co-expression module detection with WGCNA

We used the R package WGCNA [133] to detect co-expression modules based on gene expression data and a weighted correlation network. Briefly, WGNCA screens for clusters (called modules) of highly correlated genes in the provided dataset. For each gene, WGCNA also computes a "within-module connectivity", which is a measure of the number of genes from the same module it is connected with. The modules are summarized by an eigengene, which corresponds to the first principal component of the module. These eigengenes enable comparisons between modules, clustering of modules, and calculations of correlations between modules and other data, such as lipids and metabolites abundances. We provided as input the genes associated to the MFA's dimension 1 and use the within-module connectivity score to extract the top 10 most connected genes.

Modules and network representation

Network were drawn with Cytoscape 3.7.2 (https://cytoscape.org/index.html), using files extracted thanks to the exportNetworkToCytoscape function from WGCNA.

Ethics approval and consent to participate

Animals were bred at the INRAE Animal Experimental Unit PEAT (Poultry Experimental Facility, doi: 10.15454/1.5572326250887292E12; authorization C37-175-1, 2007). The experiment was conducted in compliance with the European Union Legislation and was approved by the local ethical committee in animal experimentation (Val de Loire) and by the French Ministries of Higher Education and Scientific Research, and of Agriculture and Fisheries (authorization #2873-2015112512076871).

References

1. Byerly TC, Kessler JW, Gous RM, Thomas OP. Feed Requirements for Egg Production. Poult Sci. 1980;59:2500–7.

2. Bordas A, Merat P. Correlated responses in a selection experiment on residual feed intake of adult Rhode-Island Red cocks and hens. Ann Agric Fenn. 1984;23:233–7.

3. Bordas A, Tixier-Boichard M, Merat P. Direct and correlated responses to divergent selection for residual food intake in Rhode island red laying hens. Br Poult Sci. 1992;33:741–54.

4. Swennen Q, Verhulst P-J, Collin A, Bordas A, Verbeke K, Vansant G, et al. Further Investigations on the Role of Diet-Induced Thermogenesis in the Regulation of Feed Intake in Chickens: Comparison of Adult Cockerels of Lines Selected for High or Low Residual Feed Intake. Poult Sci. 2007;86:1960–71.

5. Gabarrou J-F, Géraert P-A, Picard M, Bordas A. Diet-Induced Thermogenesis in Cockerels Is Modulated by Genetic Selection for High or Low Residual Feed Intake. J Nutr. 1997;127:2371–6.

6. Gabarrou JF, Geraert PA, Francois N, Guillaumin S, Picard M, Bordas A. Energy balance of laying hens selected on residual food consumption. Br Poult Sci. 1998;39:79–89.

7. El-Kazzi M, Bordas A, Gandemer G, Minvielle F. Divergent selection for residual food intake in Rhode Island Red egg-laying lines: Gross carcase composition, carcase adiposity and lipid contents of tissues. Br Poult Sci. 1995;36:719–28.

8. Desert C, Baéza E, Aite M, Boutin M, Le Cam A, Montfort J, et al. Multi-tissue transcriptomic study reveals the main role of liver in the chicken adaptive response to a switch in dietary energy source through the transcriptional regulation of lipogenesis. BMC Genomics. 2018;19:187.

9. Jégou M, Gondret F, Vincent A, Tréfeu C, Gilbert H, Louveau I. Whole Blood Transcriptomics Is Relevant to Identify Molecular Changes in Response to Genetic Selection for Feed Efficiency and Nutritional Status in the Pig. PENA i SUBIRÀ RN, editor. PLOS ONE. 2016;11:e0146550.

10. Louveau I, Vincent A, Tacher S, Gilbert H, Gondret F. Increased expressions of genes and proteins involved in mitochondrial oxidation and antioxidant pathway in adipose tissue of pigs selected for a low residual feed intake1. J Anim Sci. 2016;94:5042–54.

11. Gondret F, Vincent A, Houée-Bigot M, Siegel A, Lagarrigue S, Causeur D, et al. A transcriptome multi-tissue analysis identifies biological pathways and genes associated with variations in feed efficiency of growing pigs. BMC Genomics. 2017;18:244.

12. Ramayo-Caldas Y, Ballester M, Sánchez JP, González-Rodríguez O, Revilla M, Reyer H, et al. Integrative approach using liver and duodenum RNA-Seq data identifies candidate genes and pathways associated with feed efficiency in pigs. Sci Rep. 2018;8:558.

13. Horodyska J, Reyer H, Wimmers K, Trakooljul N, Lawlor PG, Hamill RM. Transcriptome analysis of adipose tissue from pigs divergent in feed efficiency reveals alteration in gene networks related to adipose growth, lipid metabolism, extracellular matrix, and immune response. Mol Genet Genomics. 2019;294:395–408.

14. Messad F, Louveau I, Koffi B, Gilbert H, Gondret F. Investigation of muscle transcriptomes using gradient boosting machine learning identifies molecular predictors of feed efficiency in growing pigs. BMC Genomics. 2019;20:659.

15. Piles M, Fernandez-Lozano C, Velasco-Galilea M, González-Rodríguez O, Sánchez JP, Torrallardona D, et al. Machine learning applied to transcriptomic data to identify genes associated with feed efficiency in pigs. Genet Sel Evol. 2019;51:10.

16. Vigors S, O'Doherty JV, Bryan K, Sweeney T. A comparative analysis of the transcriptome profiles of liver and muscle tissue in pigs divergent for feed efficiency. BMC Genomics. 2019;20:461.

17. Alexandre PA, Kogelman LJA, Santana MHA, Passarelli D, Pulz LH, Fantinato-Neto P, et al. Liver transcriptomic networks reveal main biological processes associated with feed efficiency in beef cattle. BMC Genomics. 2015;16:1073.

18. Paradis F, Yue S, Grant JR, Stothard P, Basarab JA, Fitzsimmons C. Transcriptomic analysis by RNA sequencing reveals that hepatic interferon-induced genes may be associated with feed efficiency in beef heifers1. J Anim Sci. 2015;93:3331–41.

19. Zarek CM, Lindholm-Perry AK, Kuehn LA, Freetly HC. Differential expression of genes related to gain and intake in the liver of beef cattle. BMC Res Notes. 2017;10:1.

20. Crocker Cunningham H, Cammack KM, Hales KE, Freetly HC, Lindholm-Perry AK. Differential transcript abundance in adipose tissue of mature beef cows during feed restriction and realimentation. Óvilo C, editor. PLOS ONE. 2018;13:e0194104.

21. Higgins MG, Kenny DA, Fitzsimons C, Blackshields G, Coyle S, McKenna C, et al. The effect of breed and diet type on the global transcriptome of hepatic tissue in beef cattle divergent for feed efficiency. BMC Genomics. 2019;20:525.

9. Jégou M, Gondret F, Vincent A, Tréfeu C, Gilbert H, Louveau I. Whole Blood Transcriptomics Is Relevant to Identify Molecular Changes in Response to Genetic Selection for Feed Efficiency and Nutritional Status in the Pig. PENA i SUBIRÀ RN, editor. PLOS ONE. 2016;11:e0146550.

10. Louveau I, Vincent A, Tacher S, Gilbert H, Gondret F. Increased expressions of genes and proteins involved in mitochondrial oxidation and antioxidant pathway in adipose tissue of pigs selected for a low residual feed intake1. J Anim Sci. 2016;94:5042–54.

11. Gondret F, Vincent A, Houée-Bigot M, Siegel A, Lagarrigue S, Causeur D, et al. A transcriptome multi-tissue analysis identifies biological pathways and genes associated with variations in feed efficiency of growing pigs. BMC Genomics. 2017;18:244.

12. Ramayo-Caldas Y, Ballester M, Sánchez JP, González-Rodríguez O, Revilla M, Reyer H, et al. Integrative approach using liver and duodenum RNA-Seq data identifies candidate genes and pathways associated with feed efficiency in pigs. Sci Rep. 2018;8:558.

13. Horodyska J, Reyer H, Wimmers K, Trakooljul N, Lawlor PG, Hamill RM. Transcriptome analysis of adipose tissue from pigs divergent in feed efficiency reveals alteration in gene networks related to adipose growth, lipid metabolism, extracellular matrix, and immune response. Mol Genet Genomics. 2019;294:395–408.

14. Messad F, Louveau I, Koffi B, Gilbert H, Gondret F. Investigation of muscle transcriptomes using gradient boosting machine learning identifies molecular predictors of feed efficiency in growing pigs. BMC Genomics. 2019;20:659.

15. Piles M, Fernandez-Lozano C, Velasco-Galilea M, González-Rodríguez O, Sánchez JP, Torrallardona D, et al. Machine learning applied to transcriptomic data to identify genes associated with feed efficiency in pigs. Genet Sel Evol. 2019;51:10.

16. Vigors S, O'Doherty JV, Bryan K, Sweeney T. A comparative analysis of the transcriptome profiles of liver and muscle tissue in pigs divergent for feed efficiency. BMC Genomics. 2019;20:461.

17. Alexandre PA, Kogelman LJA, Santana MHA, Passarelli D, Pulz LH, Fantinato-Neto P, et al. Liver transcriptomic networks reveal main biological processes associated with feed efficiency in beef cattle. BMC Genomics. 2015;16:1073.

18. Paradis F, Yue S, Grant JR, Stothard P, Basarab JA, Fitzsimmons C. Transcriptomic analysis by RNA sequencing reveals that hepatic interferon-induced genes may be associated with feed efficiency in beef heifers1. J Anim Sci. 2015;93:3331–41.

19. Zarek CM, Lindholm-Perry AK, Kuehn LA, Freetly HC. Differential expression of genes related to gain and intake in the liver of beef cattle. BMC Res Notes. 2017;10:1.

20. Crocker Cunningham H, Cammack KM, Hales KE, Freetly HC, Lindholm-Perry AK. Differential transcript abundance in adipose tissue of mature beef cows during feed restriction and realimentation. Óvilo C, editor. PLOS ONE. 2018;13:e0194104.

21. Higgins MG, Kenny DA, Fitzsimons C, Blackshields G, Coyle S, McKenna C, et al. The effect of breed and diet type on the global transcriptome of hepatic tissue in beef cattle divergent for feed efficiency. BMC Genomics. 2019;20:525.

35. Sugden MC, Bulmer K, Holness MJ. Fuel-sensing mechanisms integrating lipid and carbohydrate utilization. Biochem Soc Trans. 2001;29:7.

36. Resnyk CW, Carré W, Wang X, Porter TE, Simon J, Le Bihan-Duval E, et al. Transcriptional analysis of abdominal fat in genetically fat and lean chickens reveals adipokines, lipogenic genes and a link between hemostasis and leanness. BMC Genomics. 2013;14:557.

37. Miyazaki M, Dobrzyn A, Sampath H, Lee S-H, Man WC, Chu K, et al. Reduced Adiposity and Liver Steatosis by Stearoyl-CoA Desaturase Deficiency Are Independent of Peroxisome Proliferator-activated Receptor-α. J Biol Chem. 2004;279:35017–24.

38. Ropka-Molik K, Knapik J, Pieszka M, Szmatoła T. The expression of the SCD1 gene and its correlation with fattening and carcass traits in sheep. Arch Anim Breed. 2016;59:37–43.

39. Miyazaki M, Kim Y-C, Gray-Keller MP, Attie AD, Ntambi JM. The Biosynthesis of Hepatic Cholesterol Esters and Triglycerides Is Impaired in Mice with a Disruption of the Gene for Stearoyl-CoA Desaturase 1. :8.

40. Ntambi JM, Miyazaki M, Stoehr JP, Lan H, Kendziorski CM, Yandell BS, et al. Loss of stearoyl–CoA desaturase-1 function protects mice against adiposity. :5.

41. Lee S-H, Dobrzyn A, Dobrzyn P, Rahman SM, Miyazaki M, Ntambi JM. Lack of stearoyl-CoA desaturase 1 upregulates basal thermogenesis but causes hypothermia in a cold environment. J Lipid Res. 2004;45:1674–82.

42. Gabarrou J-F, Geraert PA, Williams J, Ruffier L, Rideau N. Glucose–insulin relationships and thyroid status of cockerels selected for high or low residual food consumption. Br J Nutr. 2000;83:645–51.

43. Brasaemle DL, Wolins NE. Packaging of Fat: An Evolving Model of Lipid Droplet Assembly and Expansion. J Biol Chem. 2012;287:2273–9.

44. Okumura T. Role of lipid droplet proteins in liver steatosis. J Physiol Biochem. 2011;67:629–36.

45. Carr RM, Ahima RS. Pathophysiology of lipid droplet proteins in liver diseases. Exp Cell Res. 2016;340:187–92.

46. Romeo S, Kozlitina J, Xing C, Pertsemlidis A, Cox D, Pennacchio LA, et al. Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. Nat Genet. 2008;40:1461–5.

47. Raabe M, Véniant MM, Sullivan MA, Zlot CH, Björkegren J, Nielsen LB, et al. Analysis of the role of microsomal triglyceride transfer protein in the liver of tissue-specific knockout mice. J Clin Invest. 1999;103:1287–98.

48. Karavia EA, Papachristou DJ, Liopeta K, Triantaphyllidou I-E, Dimitrakopoulos O, Kypreos KE. Apolipoprotein A-I Modulates Processes Associated with Diet-Induced Nonalcoholic Fatty Liver Disease in Mice. Mol Med. 2012;18:901–12.

49. Wang F, Kohan AB, Lo C-M, Liu M, Howles P, Tso P. Apolipoprotein A-IV: a protein intimately involved in metabolism. J Lipid Res. 2015;56:1403–18.

50. Ishihara M, Kujiraoka T, Iwasaki T, Nagano M, Takano M, Ishii J, et al. A sandwich enzyme-linked immunosorbent assay for human plasma apolipoprotein A-V concentration. J Lipid Res. 2005;46:2015–22.

51. Ito Y, Azrolan N, O'Connell A, Walsh A, Breslow J. Hypertriglyceridemia as a result of human apo CIII gene expression in transgenic mice. Science. 1990;249:790–3.

52. Wu C-L, Zhao S-P, Yu B-L. Intracellular role of exchangeable apolipoproteins in energy homeostasis, obesity and non-alcoholic fatty liver disease: Intracellular apolipoproteins and lipid metabolism. Biol Rev. 2015;90:367–76.

53. Willems van Dijk K, Rensen PC, Voshol PJ, Havekes LM. The role and mode of action of apolipoproteins CIII and AV: synergistic actors in triglyceride metabolism?: Curr Opin Lipidol. 2004;15:239–46.

54. Suganami T, Tanaka M, Ogawa Y. Adipose tissue inflammation and ectopic lipid accumulation. :9.

55. Ducheix S, Montagner A, Theodorou V, Ferrier L, Guillou H. The liver X receptor: A master regulator of the gut–liver axis and a target for non alcoholic fatty liver disease. Biochem Pharmacol. 2013;86:96–105.

56. Eberle D, Clement K, Meyre D, Sahbatou M, Vaxillaire M, Le Gall A, et al. SREBF-1 Gene Polymorphisms Are Associated With Obesity and Type 2 Diabetes in French Obese and Diabetic Cohorts. Diabetes. 2004;53:2153–7.

57. Gondret F, Ferré P, Dugail I. ADD-1/SREBP-1 is a major determinant of tissue differential lipogenic capacity in mammalian and avian species. J Lipid Res. 2001;42:8.

58. Assaf S, Hazard D, Pitel F, Morisson M, Alizadeh M, Gondret F, et al. Cloning of cDNA encoding the nuclear form of chicken sterol response element binding protein-2 (SREBP-2), chromosomal localization, and tissue expression of chicken SREBP-1 and -2 genes. Poult Sci. 2003;82:54–61.

59. Désert C, Duclos MJ, Blavy P, Lecerf F, Moreews F, Klopp C, et al. Transcriptome profiling of the feeding-to-fasting transition in chicken liver. BMC Genomics. 2008;9:611.

60. Horton JD, Goldstein JL, Brown MS. SREBPs: activators of the complete program of cholesterol and fatty acid synthesis in the liver. J Clin Invest. 2002;109:1125–31.

61. Brown MS, Goldstein JL. The SREBP Pathway: Regulation of Cholesterol Metabolism by Proteolysis of a Membrane-Bound Transcription Factor. Cell. 1997;89:331–40.

62. Nohturfft A, DeBose-Boyd RA, Scheek S, Goldstein JL, Brown MS. Sterols regulate cycling of SREBP cleavage-activating protein (SCAP) between endoplasmic reticulum and Golgi. Proc Natl Acad Sci. 1999;96:11235–40.

63. Jump DB, Oppenheimer JH. High Basal Expression and 3,5,3′-Triiodothyronine Regulation of Messenger Ribonucleic Acid $_{S14}$ in Lipogenic Tissues*. Endocrinology. 1985;117:2259–66.

64. Brown SB, Maloney M, Kinlaw WB. “Spot 14” Protein Functions at the Pretranslational Level in the Regulation of Hepatic Metabolism by Thyroid Hormone and Glucose. J Biol Chem. 1997;272:2163–6.

65. Altomonte J, Cong L, Harbaran S, Richter A, Xu J, Meseck M, et al. Foxo1 mediates insulin action on apoC-III and triglyceride metabolism. J Clin Invest. 2004;114:1493–503.

66. Kousteni S. FoxO1, the transcriptional chief of staff of energy metabolism. Bone. 2012;50:437–43.

67. Matsumoto M. Dual role of transcription factor FoxO1 in controlling hepatic insulin sensitivity and lipid metabolism. J Clin Invest. 2006;JCI27047.

68. Dong XC, Copps KD, Guo S, Li Y, Kollipara R, DePinho RA, et al. Inactivation of Hepatic Foxo1 by Insulin Signaling Is Required for Adaptive Nutrient Homeostasis and Endocrine Growth Regulation. Cell Metab. 2008;8:65–76.

69. Feng X, Jiang Y, Meltzer P, Yen PM. Thyroid Hormone Regulation of Hepatic Genes in Vivo Detected by Complementary DNA Microarray. 2000;14:9.

70. Kamagate A, Qu S, Perdomo G, Su D, Kim DH, Slusher S, et al. FoxO1 mediates insulin-dependent regulation of hepatic VLDL production in mice. :19.

71. Brown MS, Goldstein JL. Selective versus Total Insulin Resistance: A Pathogenic Paradox. Cell Metab. 2008;7:95–6.

72. James OFW, Day CP. Non-alcoholic steatohepatitis (NASH): a disease of emerging identity and importance. :7.

73. Beale E, Hammer R, Antoine B, Forest C. Disregulated glyceroneogenesis: PCK1 as a candidate diabetes and obesity gene. Trends Endocrinol Metab. 2004;15:129–35.

74. Beale EG, Harvey BJ, Forest C. PCK1 and PCK2 as candidate diabetes and obesity genes. Cell Biochem Biophys. 2007;48:89–95.

75. Herscovitz H, Ronen liana, Bilu S, Tietz A. Bile acid synthesis from HDL cholesterol and cholesterol ester by cultured chick embryo hepatocytes. Biochim Biophys Acta BBA - Lipids Lipid Metab. 1986;878:426–34.

76. Russell DW. The Enzymes, Regulation, and Genetics of Bile Acid Synthesis. Annu Rev Biochem. 2003;72:137–74.

77. Chiang JYL. Bile acids: regulation of synthesis. J Lipid Res. 2009;50:1955–66.

78. Šarenac TM, Mikov M. Bile Acid Synthesis: From Nature to the Chemical Modification and Synthesis and Their Applications as Drugs and Nutrients. Front Pharmacol. 2018;9:939.

79. Chiang JY. Regulation of bile acid synthesis. Front Biosci J Virtual Libr. 1998;3:d176-193.

80. Twisk J, Hoekman MFM, Lehmann EM, Meijer P, Mager WH, Princen HMG. Insulin suppresses bile acid synthesis in cultured rat hepatocytes by down-regulation of cholesterol 7α-hydroxylase and sterol 27-hydroxylase gene transcription. Hepatology. 1995;21:501–10.

81. Gälman C, Angelin B, Rudling M. Bile Acid Synthesis in Humans Has a Rapid Diurnal Variation That Is Asynchronous With Cholesterol Synthesis. Gastroenterology. 2005;129:1445–53.

82. Li T, Owsley E, Matozel M, Hsu P, Novak CM, Chiang JYL. Transgenic expression of cholesterol 7α-hydroxylase in the liver prevents high-fat diet-induced obesity and insulin resistance in mice. Hepatology. 2010;52:678–90.

83. Trauner M, Claudel T, Fickert P, Moustafa T, Wagner M. Bile Acids as Regulators of Hepatic Lipid and Glucose Metabolism. Dig Dis. 2010;28:220–4.

84. Lake AD, Novak P, Shipkova P, Aranibar N, Robertson D, Reily MD, et al. Decreased hepatotoxic bile acid composition and altered synthesis in progressive human nonalcoholic fatty liver disease. Toxicol Appl Pharmacol. 2013;268:132–40.

85. Krause BR, Hartman AD. Adipose tissue and cholesterol metabolism. :14.

86. Haczeyni F, Bell-Anderson KS, Farrell GC. Causes and mechanisms of adipocyte enlargement and adipose expansion: Hypertrophy and hyperplasia in adipose. Obes Rev. 2018;19:406–20.

87. Le Lay S, Krief S, Farnier C, Lefrère I, Le Liepvre X, Bazin R, et al. Cholesterol, a Cell Size-dependent Signal That Regulates Glucose Metabolism and Gene Expression in Adipocytes. J Biol Chem. 2001;276:16904–10.

88. Hirst J, King MS, Pryde KR. The production of reactive oxygen species by complex I. Biochem Soc Trans. 2008;36:976–80.

89. Curtis JM, Grimsrud PA, Wright WS, Xu X, Foncea RE, Graham DW, et al. Downregulation of Adipose Glutathione S-Transferase A4 Leads to Increased Protein Carbonylation, Oxidative Stress, and Mitochondrial Dysfunction. Diabetes. 2010;59:1132–42.

90. Maachi M, Piéroni L, Bruckert E, Jardel C, Fellahi S, Hainque B, et al. Systemic low-grade inflammation is related to both circulating and adipose tissue TNFα, leptin and IL-6 levels in obese women. Int J Obes. 2004;28:993–7.

91. Calabro P, Chang DW, Willerson JT, Yeh ETH. Release of C-Reactive Protein in Response to Inflammatory Cytokines by Human Adipocytes: Linking Obesity to Vascular Inflammation. J Am Coll Cardiol. 2005;46:1112–3.

92. Prakasam R, Fujimoto M, Takii R, Hayashida N, Takaki E, Tan K, et al. Chicken *IL-6* is a heat-shock gene. FEBS Lett. 2013;587:3541–7.

93. Territo MC, Ganz T, Selsted ME, Lehrer R. Monocyte-chemotactic activity of defensins from human neutrophils. J Clin Invest. 1989;84:2017–20.

94. Yang D, Chertov O, Bykovskaia SN, Chen Q, Buffo MJ, Shogan J, et al. β-Defensins: Linking Innate and Adaptive Immunity Through Dendritic and T Cell CCR6. Science. 1999;286:525–8.

95. Grutkoski PS, Graeber CT, Lim YP, Ayala A, Simms HH. Alpha-Defensin 1 (Human Neutrophil Protein 1) as an Antichemotactic Agent for Human Polymorphonuclear Leukocytes. :3.

96. Raj PA, Dentino AR. Current status of defensins and their role in innate and adaptive immunity. FEMS Microbiol Lett. 2002;206:9–18.

97. Rengaraj D, Truong AD, Lillehoj HS, Han JY, Hong YH. Expression and regulation of avian beta-defensin 8 protein in immune tissues and cell lines of chickens. Asian-Australas J Anim Sci. 2018;31:1516–24.

98. Kusminski CM, Scherer PE. Mitochondrial dysfunction in white adipose tissue. Trends Endocrinol Metab. 2012;23:435–43.

99. Mariman ECM, Wang P. Adipocyte extracellular matrix composition, dynamics and role in obesity. Cell Mol Life Sci. 2010;67:1277–92.

100. Henegar C, Tordjman J, Achard V, Lacasa D, Cremer I, Guerre-Millo M, et al. Adipose tissue transcriptomic signature highlights the pathological relevance of extracellular matrix in human obesity. Genome Biol. 2008;9:R14.

101. Gopal S. Syndecans in Inflammation at a Glance. Front Immunol. 2020;11:227.

102. Sun K, Kusminski CM, Scherer PE. Adipose tissue remodeling and obesity. J Clin Invest. 2011;121:2094–101.

103. Halberg N, Wernstedt-Asterholm I, Scherer PE. The Adipocyte as an Endocrine Cell. Endocrinol Metab Clin North Am. 2008;37:753–68.

104. Kukkola L, Hieta R, Kivirikko KI, Myllyharju J. Identification and Characterization of a Third Human, Rat, and Mouse Collagen Prolyl 4-Hydroxylase Isoenzyme. J Biol Chem. 2003;278:47685–93.

105. Menzies KK, Lefèvre C, Macmillan KL, Nicholas KR. Insulin regulates milk protein synthesis at multiple levels in the bovine mammary gland. Funct Integr Genomics. 2009;9:197–217.

106. Freemark M, Fleenor D, Driscoll P, Binart N, Kelly PA. Body Weight and Fat Deposition in Prolactin Receptor-Deficient Mice**This work was supported in part by grants from the NICHD (HD-24192 to M.F.), the Juvenile Diabetes Foundation (196029 to M.F.), Eli Lilly & Co. (to M.F.), and INSERM (to P.A.K.). Endocrinology. 2001;142:532–7.

107. Flint DJ, Binart N, Boumard S, Kopchick JJ, Kelly P. Developmental aspects of adipose tissue in GH receptor and prolactin receptor gene disrupted mice: site-specific effects upon proliferation, differentiation and hormone sensitivity. J Endocrinol. 2006;191:101–11.

108. Ling C, Svensson L, Odén B, Weijdegård B, Edén B, Edén S, et al. Identification of Functional Prolactin (PRL) Receptor Gene Expression: PRL Inhibits Lipoprotein Lipase Activity in Human White Adipose Tissue. J Clin Endocrinol Metab. 2003;88:1804–8.

109. Wang Y, Li J, Yan Kwok AH, Ge W, Leung FC. A novel prolactin-like protein (PRL-L) gene in chickens and zebrafish: Cloning and characterization of its tissue expression. Gen Comp Endocrinol. 2010;166:200–10.

110. Bu G, Ying Wang C, Cai G, Leung FC, Xu M, Wang H, et al. Molecular characterization of prolactin receptor (cPRLR) gene in chickens: Gene structure, tissue expression, promoter analysis, and its interaction with chicken prolactin (cPRL) and prolactin-like protein (cPRL-L). Mol Cell Endocrinol. 2013;370:149–62.

111. Yang K, Guan H, Arany E, Hill DJ, Cao X. Neuropeptide Y is produced in visceral adipose tissue and promotes proliferation of adipocyte precursor cells *via* the Y1 receptor. FASEB J. 2008;22:2452–64.

112. Rosmaninho-Salgado J, Marques AP, Estrada M, Santana M, Cortez V, Grouzmann E, et al. Dipeptidyl-peptidase-IV by cleaving neuropeptide Y induces lipid accumulation and PPAR-γ expression. Peptides. 2012;37:49–54.

113. Zhang W, Bai S, Liu D, Cline MA, Gilbert ER. Neuropeptide Y promotes adipogenesis in chicken adipose cells in vitro. Comp Biochem Physiol A Mol Integr Physiol. 2015;181:62–70.

114. Kuo LE, Kitlinska JB, Tilan JU, Li L, Baker SB, Johnson MD, et al. Neuropeptide Y acts directly in the periphery on fat tissue and mediates stress-induced obesity and metabolic syndrome. Nat Med. 2007;13:803–11.

115. Shi Y, Lin S, Castillo L, Aljanova A, Enriquez RF, Nguyen AD, et al. Peripheral-Specific Y2 Receptor Knockdown Protects Mice From High-Fat Diet–Induced Obesity. Obesity. 2011;19:2137–48.

116. Asterholm IW, Scherer PE. Enhanced Metabolic Flexibility Associated with Elevated Adiponectin Levels. Am J Pathol. 2010;176:1364–76.

117. Trujillo ME, Scherer PE. Adiponectin - journey from an adipocyte secretory protein to biomarker of the metabolic syndrome. J Intern Med. 2005;257:167–75.

118. Wang ZV, Schraw TD, Kim J-Y, Khan T, Rajala MW, Follenzi A, et al. Secretion of the Adipocyte-Specific Secretory Protein Adiponectin Critically Depends on Thiol-Mediated Protein Retention. Mol Cell Biol. 2007;27:3716–31.

119. Bordas A, Merat P. Genetic variation and phenotypic correlations of food consumption of laying hens corrected for body weight and production. Br Poult Sci. 1981;22:25–33.

120. Jehl F, Désert C, Klopp C, Brenet M, Rau A, Leroux S, et al. Chicken adaptive response to low energy diet: main role of the hypothalamic lipid metabolism revealed by a phenotypic and multi-tissue transcriptomic approach. BMC Genomics. 2019;20:1033.

121. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.

122. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

123. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.

124. R Core Team. R: A language and environment for statistical computing. R Found Stat Comput Vienna Austria [Internet]. 2017; Available from: https://www.R-project.org/

125. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.

126. Zhou X, Lindsay H, Robinson MD. Robustly detecting differential expression in RNA sequencing data using observation weights. Nucleic Acids Res. 2014;42:e91–e91.

127. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11:R25.

128. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B Methodol. 1995;57:289–300.

129. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43:D447–52.

130. Fox J, Weisberg S. An R Companion to Applied Regression [Internet]. Second. Thousand Oaks CA: Sage; 2011. Available from: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

131. Abdi H, Williams LJ, Valentin D. Multiple factor analysis: principal component analysis for multitable and multiblock data sets: Multiple factor analysis. Wiley Interdiscip Rev Comput Stat. 2013;5:149–79.

132. Lê S, Josse J, Husson F. FactoMineR: An R Package for Multivariate Analysis. J Stat Softw [Internet]. 2008 [cited 2020 May 28];25. Available from: http://www.jstatsoft.org/v25/i01/

133. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.

f) difficulté d'interprétation des résultats du sang et de l'hypothalamus

Nous n'avons présenté dans l'article qui précède que les résultats liés au tissu adipeux et au foie. Ce choix peut s'expliquer par le lien évident entre ces deux tissus et les différences entre nos deux lignées, en particulier en termes de masse de tissu adipeux ; par le fait que les données de lipidomique et de métabolomique n'étaient disponibles que pour ces deux tissus ; par la richesse des parties « Résultats » et « Discussion » tirées des analyses de ces deux tissus ; et enfin par la difficulté à interpréter les résultats liés au sang et à l'hypothalamus en relation avec les différences entre R+ et R-.

Nous avons observé 3 927 gènes DE dans l'hypothalamus et 9 028 gènes différentiellement exprimés (DE) dans le sang (pour 5 032 gènes DE dans le tissu adipeux et 6 004 dans le foie). Rappelons ici que les globules rouges des poulets (et des non-mammifères en général) sont nucléés et possèdent des mitochondries[353], et semblent être transcriptionnellement actifs[354].
L'enrichissement fonctionnel des gènes DE dans l'hypothalamus a permis de mettre en évidence 4 termes ($p_{FDR} \leq 0.05$) soutenus par 11 à 21 gènes. Ces termes de l'hypothalamus étaient liés aux Rho GTPases, impliqués dans le cytosquelette d'actine[356], l'immunité (« voie de signalisation de NF-κB » et « immunodéficience ») et le lysosome.
Dans le sang, 51 termes ont été mis en évidence ($p_{FDR} \leq 0.05$) soutenus par 9 à 199 gènes. Certains, liés à l'immunité (« voie de signalisation des récepteurs T », « cytotoxicité médié par les cellules *Natural Killer* », « système immunitaire »), sont relativement simples à relier au tissu, plus difficilement aux lignées. D'autres en revanche (citons « voie de signalisation de l'insuline ») sont intéressants et peuvent être liés aux lignées, mais plus difficilement au tissu.

La valorisation de ces résultats, en particulier ceux liés au sang, demandera, on le voit, un effort d'interprétation et éventuellement le recours à des spécialistes de ce tissu particulier composé de populations cellulaires variées.

g) travaux complémentaires sur les LNC

Nous avons utilisé les données de transcriptomique du foie et du tissu adipeux de cet article, combinées avec notre annotation étendue du génome de la poule pour étudier les LNC afin mettre en évidence des LNC ou des couples LNC:PCG potentiellement impliqués dans les différences entre nos lignées.

Nous avons mis en évidence 6 586 LNC exprimés dans le foie et 7 187 LNC exprimés dans le tissu adipeux. Parmi eux, 2 166 LNC étaient différentiellement exprimés dans le foie (958 et 1208 sur- et sous-exprimés dans les R- *versus* R+) et 1 822 (954 et 868 sur- et sous-exprimés dans les R- *versus* R+) dans le tissu adipeux.

Dans la mesure où la co-expression de deux gènes peut être due à une co-régulation, nous avons étudié le nombre de couples LNC:PCG « co-différentiellement exprimés » et l'avons comparé au nombre de couples PCG:PCG dans le même cas. L'hypothèse est ici que si les LNC ont une action régulatrice sur les PCG, nous devrions observer une plus forte proportion de couples LNC:PCG avec les deux membres DE que de couples PCG:PCG dans le même cas. Comme nous le voyons en Figure 8A ci-après, nous avons en effet observé dans les deux tissus étudiés un plus forte proportion de couples avec les deux membres DE pour les couples LNC:PCG que pour les couples PCG:PCG, quelle que soit d'ailleurs la configuration des couples en question. En effet, lorsque le LNC était DE, le PCG l'était également dans 30% à 46% des cas, contre 5% à 18% des cas pour les couples PCG:PCG. Notons cependant que pour les couples LNC:PCG en configuration « même-brin », ces pourcentages sont sans doute sur-évalués. Enfin, dans 70% des cas pour les couples LNC:PCG, les deux gènes étaient DE dans le même « sens » (tous les deux sur- ou sous-exprimés).

Nous avons ensuite mis en évidence 113 LNC ayant un *fold-change* supérieur à 6 et 100 LNC ayant un *fold-change* inférieur à 1/6 dans le foie, dont 14 et 10 étaient en couple avec un PCG également DE. Dans le tissu adipeux, nous avons mis en évidence 103 et 106 LNC ayant un *fold-change* supérieur à 6 ou inférieur à 1/6, respectivement, dont 5 et 16 en couple avec un PCG également DE. Parmi ces couples, nous avons mis en évidence le couple INRAGALG00000008498:*NDRG3*. Si ces deux gènes ne sont pas à proprement parler « corrélés » l'un avec l'autre, leurs expressions n'en présentent pas moins un motif intéressant (Figure 8B). Dans le tissu adipeux et le foie, le LNC n'est pas exprimé chez les R+ alors qu'il l'est chez les R- (TPM $\geq$ 0.2 environ). Dans le tissu adipeux, lorsque le LNC est exprimé (chez les R- donc), l'expression de *NDRG3* est plus élevée que lorsqu'il ne l'est pas (chez les R+), alors que c'est l'inverse dans le foie : lorsque le LNC n'est pas exprimé (chez les R+), l'expression de *NDRG3* est plus élevée que lorsqu'il l'est (chez les R-). Ces deux gènes sont en position divergente à 563 pb l'un de l'autre (Figure 8C)

A

Configuration	Couples℗	Adipose tissue Expr.†	one DE‡ (a)	both DE€ (b)	%(b) / (a)	p-value₳		Liver Expr.†	one DE‡ (a)	both DE€ (b)	%(b) / (a)	p-value₳	
Convergent	*LNC*	572	128	45 (30)	35.16	***		529	154	47 (35)	30.52	***	
	PCG	1095	211	27 (16)	12.80			1036	237	37 (18)	15.61		
Divergent	*LNC*	970	242	81 (57)	33.47	***		860	274	102 (70)	37.23	***	
	PCG	2774	461	25 (15)	5.42			2661	489	34 (24)	6.95		
Same-strand	*LNC*	2613	636	269 (220)	42.30	***		2257	661	304 (234)	45.99	***	
	PCG	2894	528	63 (42)	11.93			2750	600	88 (59)	14.67		
Antisense	*LNC*	764	218	67 (37)	30.73	***		696	265	107 (73)	40.38	***	
	PCG	619	117	12 (8)	10.26			594	147	27 (13)	18.37		
Sense of introns	*LNC*	75	14	4 (3)	28.57	*n.s*		61	15	6 (6)	40.00	*n.s*	
	PCG	168	38	6 (4)	15.79			529	43	22 (13)	51.16		

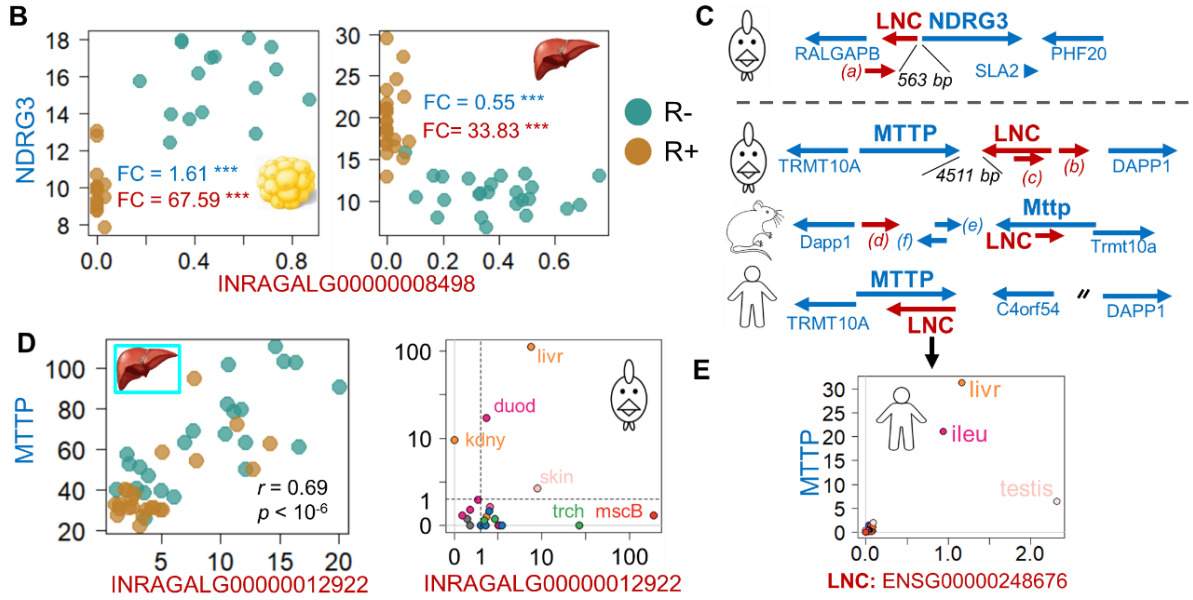


Figure 8 | Expression des couples LNC:PCG. (**A**) Vue générale de l'expression des couples LNC:PCG et PCG:PCG dans le tissu adipeux et le foie, et comparaison des pourcentages de couples LNC:PCG et PCG:PCG dont les deux membres sont différentiellement exprimés. Entre parenthèse est indiqué le nombre de *fold-change* de même « sens ». (**B**) Co-expression de INRAGALG00000009498:*NDRG3* dans le tissu adipeux (gauche) et dans le foie (droite), en RPKM. (**C**) Configuration génomique de *NDRG3* et de son LNC chez la poule (au-dessus de la ligne pointillée), pour lesquels aucune synténie n'a été trouvée chez l'humain ni la souris, et de *MTTP* et de son LNC ou du potentiel LNC dans 3 espèces (en-dessous de la ligne pointillée) : poule (en haut), souris (au milieu) et humain (en bas). Les échelles ne sont pas respectées, et la distance entre *C4orf54* et *DAPP1* chez l'humain est plus importante que suggérée. (a) ENSGALG00000029930, (b) INRAGALG00000012927, (c) INRAGALG00000012924, (d) *Gm19708*, (e) ENSMUSG00000090066, (f) *Gm5105*. (**D**) Co-expression entre *MTTP* et INRAGALG00000012922, tous les deux membres du module WGCNA "turquoise", en utilisant les données d'expression de ce travail (en RPKM, à gauche), du travail réalisé dans le cadre de l'article 1 de la présente thèse (en TPM, à droite). (**E**) Corrélation d'expression entre *MTTP* et son potentiel LNC chez l'humain, en TPM. *n.s* : non-significatif, * : p ≤ 0.05, *** : p ≤ 0.001, ℗ : composition des couples. LNC : couples LNC:PCG, PCG : couples PCG:PCG. † : nombre de couples exprimés, ‡ : nombre de couples avec le LNC DE (première ligne de chaque configuration) ou l'un des deux PCG DE (seconde ligne), € : nombre de couples avec les deux gènes DE, ₳ : *p-value* d'un test du χ^2 comparant la proportion du nombre de couples avec les deux membres DE versus le nombre de couples avec au moins un membre DE, FC : *fold-change* entre R- and R+. "duod" : duodénum, "ileu" : iléon, "kdny" : rein, "mscB" : muscle pectoral, "skin" : peau, "testis" : testicule, "trch" : trachée.

Ensuite, nous nous sommes intéressés aux LNC présents dans les modules formés par WGCNA. Nous avons mis en évidence 13 couples LNC:PCG (5 chez les R- et 8 chez les R+) dont les deux membres étaient présents dans un module, indiquant (*i*) qu'ils présentaient une forte corrélation d'expression et (*ii*) que les deux gènes participaient à la séparation des deux lignées. Le couple en position convergente LNC:*MTTP*, présent dans le module enrichi en termes associés à la voie de signalisation des PPAR, est d'un intérêt particulier. En effet, *MTTP* est impliqué dans l'assemblage des lipoprotéines, qui transportent les acides gras du foie vers les tissus cibles, notamment le tissu adipeux. Le profil d'expression de ces deux gènes est présenté Figure 8D (à gauche). Nous avons identifié d'autres tissus dans lesquels les deux gènes sont exprimés en utilisant les données de l'article 1 de la présente thèse (Figure 8D). Tous deux sont exprimés dans le foie, et dans une moindre mesure le duodénum, le rein, la peau, la trachée et le muscle pectoral. Ce couple est conservé entre poule et mammifères en termes de position (Figure 8C). En utilisant des données GTEx, nous avons donc étudié chez l'humain la corrélation d'expression entre *MTTP* et deux gènes qui pourraient être des orthologues du LNC chez la poule (Figure 8E). Ces deux gènes semblent être exprimés concomitamment dans le foie, l'iléon, le muscle squelettique et le testicule.

Ces résultats relatifs à la proportion de couples LNC:PCG dont les deux membres sont DE suggèrent que la seule proximité des deux gènes suffit à ce qu'ils soient tous deux DE, indépendamment de leur classification. Les mécanismes sous-jacents à cette « association » restent en revanche à explorer. On peut imaginer des mécanismes spécifiques de certaines configurations, ou bien des mécanismes génériques : sont-ils sous le contrôle d'un régulation commune (provoquant une transcription bidirectionnelle dans le cas des couples divergents), le transcrit LNC généré à proximité du gène PCG permet-il le recrutement d'éléments nécessaires à l'expression du PCG (majoritairement par une régulation positive d'ailleurs), ou bien la transcription du LNC est-elle nécessaire à celle du PCG sans que le transcrit LNC ne joue un rôle particulier dans cette transcription (voir à ce propos l'exemple de *Bend4* et *Bendr* développé dans l'introduction) ?
Enfin, les rôles éventuels du LNC en couple avec *MTTP* sur la formation des lipoprotéines et du LNC en couple avec *NDRG3* (dont les rôles sont d'ailleurs mal connus[357]) sur son expression restent à élucider.

2. Réponse adaptative de la poule à un régime hypo-énergétique : le rôle clef du métabolisme hypothalamique des lipides mis en évidence par une approche associant phénotypes et transcriptomes multi-tissus (article 4)

a) contexte

À l'échelle mondiale, la filière œufs de poule utilise des poules pondeuses sélectionnées par l'une des quelques entreprises du secteur. Les lignées desquelles sont issus ces animaux sont sélectionnées dans des conditions contrôlées, mais les animaux de production sont ensuite utilisés dans le monde entier, et sont donc exposés à des conditions environnementales très diverses. Ainsi, les animaux peuvent subir des variations de température, de composition de la ration alimentaire, ou bien de contenu énergétique. C'est sur ce dernier point que se concentre le présent article. Il s'agit ici de mieux comprendre les gènes et processus biologiques impliqués dans la réponse adaptative de poules pondeuses à un régime hypo-énergétique par une approche transcriptomique sur 4 tissus. En effet, si quelques études se sont intéressées aux effets d'un régime hypo-énergétique sur les performances de poules pondeuses, aucune n'a investigué les mécanismes sous-jacents aux variations de performances au niveau transcriptomique. Afin de mettre en évidence une éventuelle interaction entre la réponse au régime hypo-énergétique et l'efficience alimentaire, nous avons utilisé des animaux de deux lignées divergentes pour ce caractère : les R+, la lignée inefficiente et les R-, la lignée efficiente.

b) matériels et démarche

Les animaux de ces deux lignées ont été nourris *ad libitum* avec un régime standard ou un régime hypo-énergétique (15% d'énergie en moins par rapport au standard) pendant 14 semaines (entre 17 et 31 semaines d'âge). Cette ration hypo-énergétique mime celle à laquelle des poules pondeuses pourraient être exposées dans des pays dans lesquels l'accès aux ingrédients énergétiques serait difficile voire impossible.

Pour étudier les effets de ces régimes sur les animaux, deux types de données ont été collectées. D'une part, des données relatives aux performances des animaux et à leur prise alimentaire : poids du corps, intensité de ponte (nombre d'œufs pondus par jour sur une période donnée, divisé par le nombre de poules présentes), nombre, poids et solidité des œufs pondus, etc., sur 96 animaux au total : 45 R+ (34 du régime contrôle et 11 du régime hypo-énergétique) et 51 R- (36 du régime contrôle et 15 du régime hypo-énergétique). D'autre part, à l'issue de la période

de stress, à 31 semaines d'âge, des échantillons de quatre tissus ont été prélevés sur 32 animaux (8 R+ et 8 R- de chaque régime) afin d'en analyser le transcriptome. Ces quatre tissus étaient le tissu adipeux, lieu du stockage et de la mobilisation des acides-gras qui composent les réserves énergétiques des animaux, le foie, lieu notamment de la synthèse des lipides dont les acides-gras et le cholestérol, l'hypothalamus, centre de régulation de l'homéostasie énergétique, et enfin le sang, tissu circulant qui assure le transport des nutriments et messagers moléculaires dans tout l'organisme.

Les données relatives aux performances des animaux et à leur prise alimentaire ont été analysées en utilisant un modèle d'analyse de variance prenant en compte la lignée, le régime et l'interaction lignée × régime. Les analyses de différentiel d'expression ont été réalisées à l'aide du package R « edgeR », à l'aide d'un modèle prenant là encore en compte la lignée, la régime et l'interaction lignée × régime.

c) résultats

Nos résultats ont montré tout d'abord que la différence de contenu énergétique de la ration n'avait pas affecté les performances des animaux, mais avait en revanche, dans le groupe nourri avec le régime hypo-énergétique, entrainé une augmentation de leur prise alimentaire et une diminution de leur masse de gras corporel (voir Table 1 de l'article). Dit autrement, les animaux nourris avec l'aliment appauvri en énergie semblent avoir compensé cet appauvrissement en consommant plus et en mobilisant leurs réserves corporelles. Ces résultats étaient les mêmes quelle que soit la lignée des animaux, montrant une absence d'interaction entre la ration et la lignée à ce niveau.

Ensuite, au niveau des transcriptomes tissulaires, nous avons observé un contraste important entre, d'un côté, le foie et le tissu adipeux, qui ne semblent pas avoir été affectés du point de vue transcriptomique : ils totalisent moins de 20 gènes différentiellement exprimés à eux deux. En revanche, le sang et surtout l'hypothalamus ont été fortement impactés avec respectivement 1 334 et 2 700 gènes différentiellement exprimés entre les deux régimes. Encore une fois, aucune interaction n'a été observée entre la ration et la lignée.

Dans le sang, les gènes différentiellement exprimés étaient liés à des fonctions telles que la synthèse des acides-aminés ou du cholestérol (Table 3), et semblent indiquer un rôle potentiel du sang dans l'adaptation à la ration appauvrie en énergie.

Dans l'hypothalamus, la ration appauvrie semble avoir eu plusieurs effets (Table 2). D'abord, nous avons observé une possible modification de l'organisation synaptique, avec notamment l'activation de la synthèse protéique, potentiellement alimentée en énergie par la β-oxydation

des acides gras, ou encore l'altération du métabolisme du cholestérol (Figures 2 et 5 de l'article). Nous avons également mis en évidence une voie métabolique pouvant expliquer la hausse de la prise alimentaire, à travers l'activation de la voie de endocannabinoïdes, des molécules connues pour réguler la prise alimentaire, et dont la synthèse d'un des précurseurs, l'acide arachidonique, un acide gras poly-insaturé, semble être activée dans l'hypothalamus des animaux nourris avec la ration hypo-énergétique. Nous avons en effet observé une expression différentielle de différents gènes de ces voies (Figure 6 de l'article).

d) discussion et conclusion

Ce travail a donc permis d'explorer les effets d'une ration appauvrie en énergie sur les performances, la prise alimentaire et les transcriptomes tissulaires des animaux des deux lignées R+ et R-. Malgré le faible effectif des animaux par niveau de facteurs (n = 8 en transcriptomique et n entre 11 et 36 pour les performances), nous avons pu observer entre les animaux des deux régimes des variations suffisamment fortes pour être significatives, aussi bien au niveau des performances que des transcriptomes. En revanche aucune des variables étudiées ne présentait d'interaction significative entre régime et efficience alimentaire. En effet, bien qu'une tendance ait pu être observée, elle n'était pas suffisamment forte pour être significative avec de tels effectifs. Ainsi les R- du groupe hypo-énergétique ont eu tendance à voir leur poids du corps diminuer dans des proportions plus élevées que les R+ du même groupe (-8% environ versus -1%). Les résultats majeurs ont été une réaction adaptative des animaux au régime *via* une mobilisation des réserves corporelles et une hausse de la prise alimentaire, en accord avec la bibliographie[358–360]. Concernant cette dernière observation, nous avons mis en évidence dans l'hypothalamus un mécanisme impliquant les endocannabinoïdes qui peut contribuer à expliquer cette hausse de prise alimentaire. Si les endocannabinoïdes sont connus pour leur implication dans la régulation de la prise alimentaire[361,362], ce travail est le premier à mettre en évidence un tel mécanisme dans le cadre de l'adaptation à un régime hypo-énergétique.

e) article publié[363]

Cet article a fait l'objet :

- d'une publication dans *BMC Genomics* : **Jehl, F.**, Désert, C., Klopp, C. et al. Chicken adaptive response to low energy diet: main role of the hypothalamic lipid metabolism revealed by a phenotypic and multi-tissue transcriptomic approach. BMC Genomics 20, 1033 (2019). https://doi.org/10.1186/s12864-019-6384-8, reproduite ci-après ;

- d'une communication orale : **F. Jehl**, M. Brenet, A. Rau, C. Désert, M. Boutin, S. Leroux, D. Esquerré, C. Klopp, D. Gourichon, A. Collin, F. Pitel, T. Zerjal and S. Lagarrigue. Layers response to suboptimal diet through phenotypic and transcriptomic changes in four tissues. 69th Annual Meeting of the European Federation of Animal Science,
  Dubrovnik (Croatie), le 30 août 2018 ;

- d'un poster : **F. Jehl**, C. Klopp, M. Brenet, A. Rau, C. Désert, M. Boutin, S. Leroux, K. Muret, D. Esquerré, D. Gourichon, T. Burlot, Anne Collin, F. Pitel, T. Zerjal, S. Lagarrigue. Phenotype and multi-tissue transcriptome response to diet-energy changes in laying hens. International Plant & Animal Genome XXVII (PAG),
  San Diego (États-Unis), le 14 janvier 2019.

Jehl *et al. BMC Genomics* (2019) 20:1033
https://doi.org/10.1186/s12864-019-6384-8

BMC Genomics

RESEARCH ARTICLE Open Access

Chicken adaptive response to low energy diet: main role of the hypothalamic lipid metabolism revealed by a phenotypic and multi-tissue transcriptomic approach

F. Jehl[1], C. Désert[1], C. Klopp[2], M. Brenet[1], A. Rau[3], S. Leroux[4], M. Boutin[1], L. Lagoutte[1], K. Muret[1], Y. Blum[5], D. Esquerré[6], D. Gourichon[7], T. Burlot[8], A. Collin[9], F. Pitel[4], A. Benani[10], T. Zerjal[2*] and S. Lagarrigue[1*]

Abstract

Background: Production conditions of layer chicken can vary in terms of temperature or diet energy content compared to the controlled environment where pure-bred selection is undertaken. The aim of this study was to better understand the long-term effects of a 15%-energy depleted diet on egg-production, energy homeostasis and metabolism via a multi-tissue transcriptomic analysis. Study was designed to compare effects of the nutritional intervention in two layer chicken lines divergently selected for residual feed intake.

Results: Chicken adapted to the diet in terms of production by significantly increasing their feed intake and decreasing their body weight and body fat composition, while their egg production was unchanged. No significant interaction was observed between diet and line for the production traits. The low energy diet had no effect on adipose tissue and liver transcriptomes. By contrast, the nutritional challenge affected the blood transcriptome and, more severely, the hypothalamus transcriptome which displayed 2700 differentially expressed genes. In this tissue, the low-energy diet lead to an over-expression of genes related to endocannabinoid signaling (*CN1R*, *NAPE-PLD*) and to the complement system, a part of the immune system, both known to regulate feed intake. Both mechanisms are associated to genes related polyunsaturated fatty acids synthesis (*FADS1*, *ELOVL5* and *FADS2*), like the arachidonic acid, a precursor of anandamide, a key endocannabinoid, and of prostaglandins, that mediate the regulatory effects of the complement system. A possible regulatory role of *NR1H3* (alias *LXRα*) has been associated to these transcriptional changes. The low-energy diet further affected brain plasticity-related genes involved in the cholesterol synthesis and in the synaptic activity, revealing a link between nutrition and brain plasticity. It upregulated genes related to protein synthesis, mitochondrial oxidative phosphorylation and fatty acid oxidation in the hypothalamus, suggesting reorganization in nutrient utilization and biological synthesis in this brain area.

Conclusions: We observed a complex transcriptome modulation in the hypothalamus of chicken in response to low-energy diet suggesting numerous changes in synaptic plasticity, endocannabinoid regulation, neurotransmission, lipid metabolism, mitochondrial activity and protein synthesis. This global transcriptomic reprogramming could explain the adaptive behavioral response (i.e. increase of feed intake) of the animals to the low-energy content of the diet.

Keywords: Transcriptome, Lipid, Feed intake, Adaptation, Hypothalamus, Chicken

* Correspondence: tatiana.zerjal@inra.fr; sandrine.lagarrigue@agrocampus-ouest.fr
[2]SIGENAE Plateform, INRA, 31326 Castanet-Tolosan, France
[1]PEGASE UMR 1348, INRA, AGROCAMPUS OUEST, 35590 Saint-Gilles, France
Full list of author information is available at the end of the article

Background

The egg-production sector uses genetically selected chicken breeds bought from a few breeding companies. While the purebred selection process usually takes place in a controlled environment, commercial layers are exposed to a wide diversity of environments, some being more challenging than others because of stressors like high heat, sub-optimal diet composition or low diet energy content. In this study we investigated, on laying hens, the effects that a 15%-energy depleted diet provided ad libitum over a long period (14 weeks) has on the transcriptome of several energy-related tissues to verify if animal performance changes related to the low energy intake were due to underlying mechanisms at the transcriptomic level. The low-energy diet used in this study resembles the type of diet that can used for layer production in countries where, for diverse reasons, access to protein or oil happens to be too costly or impossible due to the lack of supply. While several studies have investigated the effect of a low-energy diet on the performances of laying hens, no study has analyzed the tissue mechanisms underlying performance variations at the transcriptomic level. As examples, Grobas et al. [1] observed an increase in feed intake, a decrease in body weight gain and no difference in egg production rate and egg weight in layers fed ad libitum a 2680 kcal/kg diet compared to a 2810 kcal/kg diet, both with the same protein content levels per kilocalorie of energy, from 22 to 65 weeks of age. Harms et al. [2] observed the same results regarding feed intake, body weight gain, egg production rate and egg weight for layers fed a 2519 kcal/kg diet compared to a 2798 kcal/kg control diet from 36 to 44 weeks of age, with adjusted levels of amino-acids. On the contrary, Murugesan and Persia [3] observed no effects on egg production, body weight and feed intake, but only a reduction of the abdominal fat pad mass in layers fed ad libitum a 2790 kcal/kg diet, compared to a 2880 kcal/kg control diet, both diets having approximately the same crude proteins content, from 28 to 39 weeks In this context, we investigated the effects of a low-energy diet on the performances and feed intake together with the transcriptomes of four tissue of adult layers fed ad libitum two diets differing in energy content (2321 kcal/kg for the low-energy diet versus 2710 kcal/kg for the commercial diet) from 17 to 31 weeks of age. Since feed efficiency is a key factor for energy allocation and is a trait of economic importance, we hypothesized a possible interaction between feed efficiency and the response to the energy-depleted diet. We therefore compared the response to the low-energy diet between two brown egg layer lines divergently selected for the residual feed intake (RFI) [4] to evaluate such a potential interaction between diet and feed efficiency factors. The RFI is the difference between the predicted feed intake estimated considering body weight and egg production, and the observed feed intake. The four tissues used to explore the transcriptomic mechanisms at work in response to the low-energy diet on the same animals as those used for the performance analysis were the liver, the adipose tissue, the blood and the hypothalamus, all related to energy homeostasis. The adipose tissue is crucial for fatty acid storage, the main form of energy storage, and mobilization. The liver is a key organ for lipogenesis in birds [5], in addition to many other physiological processes such as oxidation, secretion and detoxification. The hypothalamus is an important center for the regulation of feed intake, and blood is a circulating tissue that gathers and transports nutrients, hormones, proteins and cell waste throughout an organism. To the best of our knowledge, such a study analyzing both laying performances and four tissue transcriptomes in response to an energy-depleted diet has not yet been undertaken in layers.

Results

Diet energy change had little effect on production traits but affected feed intake and body composition

The line, diet and interaction effects on body weight, egg production and shell strength, feed intake (FI), residual feed intake (RFI) and abdominal adipose weight after 14 weeks of the low-energy diet are summarized in Table 1. The diet energy content difference had no effect on egg production, i.e. on laying rate, egg weight and egg mass. In contrast, we observed a significant decrease in body weight at 31 weeks (on average for both lines, − 4.4%, $p < 0.05$) in the LE group compared to the CT group, despite the fact that at the beginning of the trial (17 weeks of age), the LE group was slightly heavier than the CT group (on average, + 3%, $p < 0.05$, Additional file 1). We also observed a significant ($p < 0.05$) increase of feed intake in the LE group over 28 to 31th week of age, without significant interaction with the line ($p = 0.50$). It can however be noted that the increase in feed intake in response to the LE diet is smaller in the R- line (+ 145 g) than in the R+ line (+ 270 g), which can be related to the fact that the R- line generally eats less; the interaction between diet and line remains however not significant. The calculated RFI was significantly higher in the LE group, meaning that the animals were less feed efficient than the CT group. Finally, the LE group had at 31 weeks of age a significantly lower ratio of abdominal adipose tissue weight to body weight compared to the CT group (on average, − 0.72, $p < 0.05$), even if the body weight significantly decreased at the same age (on average − 4.4%, $p < 0.05$) indicating a higher decrease of abdominal tissue (on average, − 20.6%, $p < 0.05$). Concerning the line factor, as expected, we observed significant differences on FI, RFI and abdominal adipose

Table 1 Means (±SD) and significance for production, feed efficiency and body composition traits, for the effect of the diet, the line and their interaction

	{R+,CT}[a]	{R+,LE}[a]	{R-,CT}[a]	{R-,LE}[a]	Diet[b]	Line[b]	Diet × Line[b]
Body weight, week 31 (g)	2162.35 (±165.33)	2142.46 (±129.28)	2089.44 (±216.87)	1925.40 (±217.32)	*	**	0.11
Laying intensity (%)	86.17 (±11.92)	87.73 (±7.81)	86.87 (±5.44)	84.59 (±8.58)	0.70	0.50	0.54
Egg number	60.94 (±9.33)	62.18 (±9.93)	61.17 (±6.16)	60.47 (±7.43)	0.93	0.86	0.60
Egg weight (g)	47.91 (±3.11)	46.80 (±2.98)	48.08 (±2.25)	47.61 (±1.82)	0.21	0.53	0.60
Egg mass (g)[c]	1166.41 (±181.31)	1182.36 (±210.53)	1118.36 (±108.85)	1055.80 (±126.99)	0.43	*	0.27
Static stiffness ($N.mm^{-1}$)	109.68 (±18.75)	104.64 (±15.58)	126.75 (±18.39)	118.95 (±18.76)	0.12	***	0.75
Feed intake (g)[c]	4128.47 (±426.94)	4398.10 (±551.14)	2583.92 (±308.26)	2728.73 (±419.65)	*	***	0.50
Energy intake (kcal)[c]	11,188.16 (±1157.00)	10,207.97 (±1279.19)	7002.41 (±835.38)	6333.39 (±974.01)	**	***	0.52
RFI ($g/21d^{-1}$)[c]	868.36 (±329.66)	1152.32 (±390.52)	− 614.35 (±134.93)	− 196.81 (±211.78)	***	***	0.28
Abdominal adipose weight at 31 weeks (g)	73.33 (±21.10)	57.10 (±18.61)	129.83 (±44.23)	105.00 (±31.67)	*	***	0.64
Ratio of abdominal adipose weight to body weight at 31 weeks (%)	3.37 (±0.83)	2.65 (±0.78)	5.96 (±1.39)	5.24 (±1.09)	*	***	1

[a]Values represent the line/treatment group means for each trait (±standard deviation). R+ refers to low feed efficient layers and R- to high feed efficient layers, CT to control group and LE to low energy diet. The number of animals analyzed are: R+,CT $n = 34$, R+,LE $n = 11$, R-,CT $n = 36$, R-,LE $n = 15$
[b]***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$
[c]Feed-related traits were measured between 28 and 31 weeks of age

weight. The significant line effect for the body weight at 31 weeks, for which the interaction *p*-value was the lowest and close to 0.10 is due to the {R-,LE} group, the animals of which are lighter than in the three other groups. However, we observed no significant differences between the body weight of R+ and R- from the control group, as expected since the divergent selection on RFI was performed at constant body weight. Both lines, regardless of their RFI, reacted in a similar way to the energy-depleted diet by increasing their feed intake. However, this increase in feed ingestion was not sufficient to avoid body weight loss in the R- fed with the LE diet and depletion of the energy reserves (body fat). To explore the molecular mechanisms underlying this adaptation, we analyzed gene expression of several tissues of birds from these two lines and diets.

Diet energy change leads to transcriptomic modifications, mainly in hypothalamus and blood

To explore the genes involved in the response of birds to the two diets, we analyzed the transcriptomic changes associated with diet changes in the adipose tissue, blood, hypothalamus and liver. A total of 16,461 genes were expressed in at least one of the four tissues considered, and represents 66% of the 24,881 genes from Ensembl v93 annotation (Fig. 1a and b). Of these 16,461 genes, 13,567, 11,440, 15,307 and 12,873 were expressed in the adipose, blood, hypothalamus and liver, respectively (Fig. 1b), and 10,314 (41%) were expressed in all four tissues (Fig. 1a). Some of these genes were tissue-specific, representing 1.34% (adipose) to 10.8% (hypothalamus) of the total number of genes expressed in the tissues (Fig. 1a, Additional file 2). The hypothalamus had markedly higher gene-specificity, with 1653 genes expressed only in this tissue. It also had the greatest number of total expressed genes (15307). Strikingly, diet change had a large effect on the hypothalamic and blood transcriptomes, with 2700 and 1334 differentially expressed genes (DEG), respectively, while the hepatic and adipose tissue transcriptomes were almost unaffected (15 and 2 DEG, respectively) (Fig. 1c and d, Additional file 3). The line had a major effect in all tissues, with 3143, 4631, 1874 and 2480 DEG in the adipose, blood, hypothalamus and liver, respectively. As only a very small number of significant interactions ($p_{FDR} < 0.05$) were observed (Fig. 1c), allowing for an independent analysis of the line and diet factors, the present paper focuses only on the diet effect.

Functional characterization of hypothalamic transcriptome changes upon diet energy challenge

Among the 2700 DEG detected in the hypothalamus in response to the diet energy change, 1438 and 1262 genes were over- and under-expressed, respectively, in the LE group compared to the control. We characterized these two DEG lists using KEGG pathway term enrichment as described in Methods. For the over- and under-expressed gene lists, 26 and 44 pathways ($p_{FDR} < 0.05$) were significantly enriched (Additional file 4). The 10 top terms with the lowest p_{FDR} for both DEG lists are presented in Table 2.

Pathways associated with the under-expressed genes (Table 2A) comprised 91 under-expressed genes related to different types of synapses: glutamatergic, dopaminergic

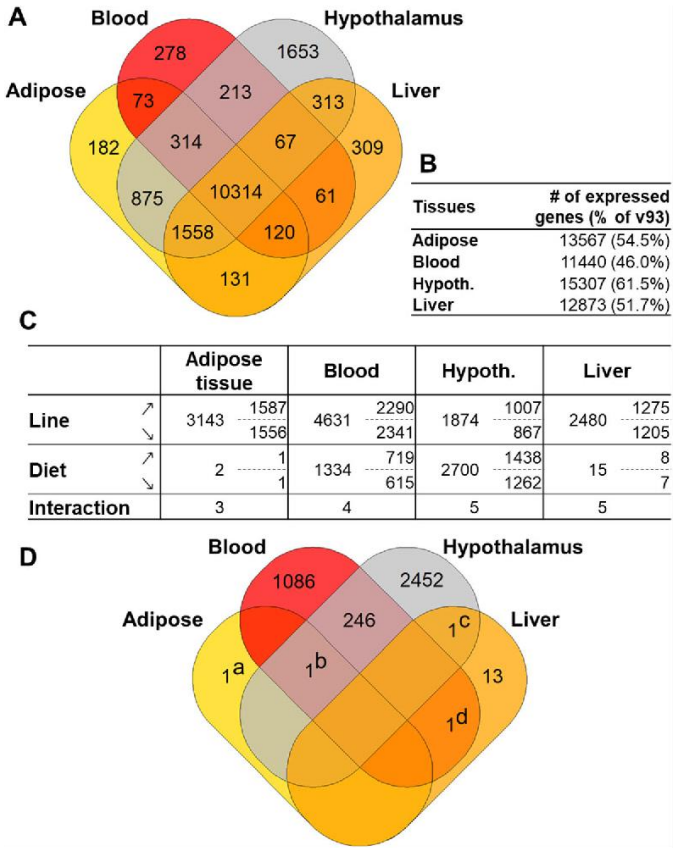


Fig. 1 Overview of gene expression and differential expression between diets in the adipose tissue, blood, hypothalamus and liver. **a** Venn diagram of the genes expressed and shared in the four tissues. **b** Total number of genes expressed in each tissue; between brackets, percentage of v87 annotation (24,881 genes). **c** Differentially expressed genes (DEG) in each tissue (columns) and each factors, Line, Diet and Interaction (rows). The total number of DEG (left) and the details of the number of up- (↗) and down-expressed genes (↘) in LE diet (or R+ line) compared to CT (to R- line) are indicated. Hypoth.: Hypothalamus. **d** Venn diagram of the DEG between diets in the four tissues. Single genes in the diagram are: (a) ENSGALG00000002503 (SFTPA2) (b) ENSGALG00000031497 (no HGNC), (c) ENSGALG00000026507 (FDX1) and (d) ENSGALG00000006099 (ZFPM1)

and GABAergic synapses, as well as the synaptic vesicle cycle or axon guidance. Among these genes were notably *GRIA1*, *GRIA3* and *GRIA4* that code for subunits of the glutamate receptor, the predominant excitatory neurotransmitter in the nervous system; *DDC*, that code for an enzyme involved in the synthesis of dopamine, a neurotransmitter involved in the reward system, and *DRD3* that code for a subunit of the dopamine receptor; *GABRQ*, *GABRG2*, *GABRR2* that code for subunits of the receptor to the gamma-aminobutyric acid (GABA), the major inhibitory neurotransmitter.

Pathways associated with over-expressed genes in LE compared to CT (Table 2B) were related to the "Ribosome" and several metabolic pathways. "Ribosome" comprises 83 ribosomal Protein genes, of which 41 Ribosomal Protein L (*RPLx*) genes, 27 Ribosomal Protein S (*RPSx*), as well as 8 Mitochondrial Ribosomal Protein L (*MRPLx*) and 5 Mitochondrial Ribosomal Protein S (*MRPSx*). Among the metabolic pathways, energy-related pathways appear to be most affected. Indeed, we found an over-representation of genes associated with oxidative phosphorylation, a process that involves a series of oxidation-reduction reactions in mitochondria, resulting in the phosphorylation of ADP to produce ATP. Among these genes, 31 were related to one of the 5 protein complexes constituting the respiratory chain located in the inner mitochondrial membrane: 15 genes for the complex I (NADH:ubiquinone oxidoreductase), 8 genes the complex II (succinate:ubiquinone oxidoreductase), 3 genes for the complex III (ubiquinol:ferricytochrome C oxidoreductase), 2 genes for the complex IV (cytochrome C

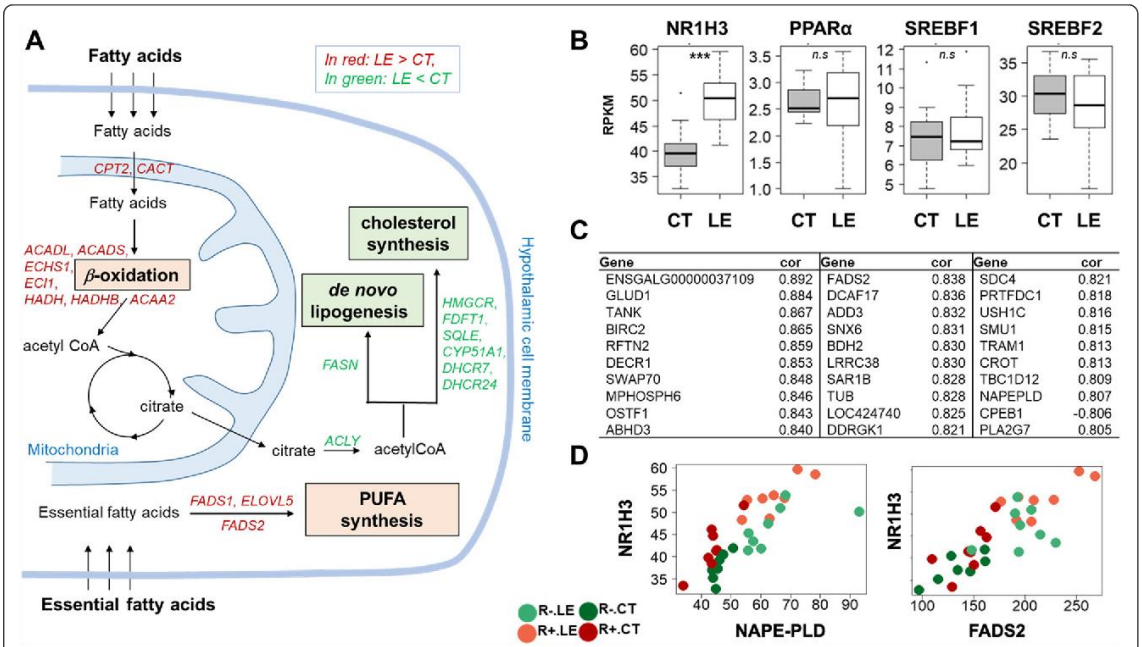
Table 2 Top 10 (based on p_{FDR}) KEGG pathways associated with under-expressed (**A**) and over-expressed DEG (B) in the hypothalamus

Term	# of genes	p_{FDR}
A. Under-expressed genes in LE compared to CT		
Synaptic vesicle cycle	22	7.36×10^{-11}
Glutamatergic synapse	26	1.79×10^{-08}
Dopaminergic synapse	26	2.37×10^{-07}
Axon guidance	25	5.62×10^{-07}
Oxytocin signaling pathway	27	2.46×10^{-06}
Circadian entrainment	20	2.50×10^{-06}
Oocyte meiosis	21	7.03×10^{-06}
Protein processing in endoplasmic reticulum	26	2.04×10^{-05}
Nicotine addiction	12	2.04×10^{-05}
GABAergic synapse	17	5.18×10^{-05}
B. Over-expressed genes in LE compared to CT		
Ribosome	83	1.03×10^{-67}
Metabolic pathways	166	2.57×10^{-25}
Oxidative phosphorylation	46	3.26×10^{-22}
Glycine, serine and threonine metabolism	15	7.73×10^{-08}
Fatty acid metabolism	15	1.81×10^{-06}
Fatty acid degradation	14	2.52×10^{-06}
Valine, leucine and isoleucine degradation	14	3.18×10^{-06}
PPAR signaling pathway	16	3.65×10^{-05}
Carbon metabolism	19	1.54×10^{-04}
Alanine, aspartate and glutamate metabolism	10	4.70×10^{-04}

oxidase) and 2 genes for the complex V (FoF1-ATP synthetase), in addition to SLC25A4, the ADP/ATP translocase 1. More than 21 of them are located in the mitochondrial genome. In addition, genes involved in fatty acid transport (*ACSBG1*, *APOA1*, *APOC3*, *DBI*, *SLC27A1*, *FABP4*, *FABP7*, *SCP2*), the fatty acid β-oxidation in the mitochondria (*CPT2*, *CACT*, *ACADL*, *ACADS*, *ECHS1*, *ECI1*, *HADH*, *HADHB*, *ACAA2*), and to a lesser extent, in the peroxisomes (*ACAA1*, *ACOX*, *ECI2*) were also over-expressed. On the contrary, genes involved in the de novo lipogenesis were significantly under-expressed, in particular *FASN*, that codes for a key enzyme of the saturated fatty acid synthesis, and *ACLY* that codes for the primary enzyme involved in the synthesis of cytosolic acetyl-CoA from citrate. Similarly genes involved in the cholesterol synthesis such as *HMGCR*, *FDFT1*, *SQLE*, *CYP51A1*, *DHCR7*, and *DHCR24* were also under-expressed. Interestingly, we observed an over-expression of genes involved in the biosynthesis of ω3 and ω6 polyunsaturated fatty acids, with *FADS2*, *ELOVL5*, *FADS1*, *ELOVL2* and (see top 5 and 19 KEGG term). It is noteworthy that one of the products of this pathway, the arachidonic acid, can be used by the enzyme coded by *NAPEPLD*, which is over-expressed (FC = 1.93, $pFDR = 6.86 \times 10^{-11}$) as a substrate for the synthesis of anandamide. Since the lipid metabolism was largely impacted (Fig. 2a), we studied the transcription factors related to this metabolism (Fig. 2b). The expressions of *PPARA*, *SREBF2* and *SREBF1* genes were not affected (FC = 1; 0.88 and 1.08 respectively, with $p_{FDR} = 0.99$; 0.44 and 0.79, respectively). On the other hand, *NR1H3* (alias *LXRA*) was significantly over-expressed (FC = 1.55, $p_{FDR} = 2 \times 10^{-6}$). The 30 genes most correlated ($r > 0.8$) to *NR1H3* are showed in Fig. 2c in which can be found *FADS2* and *NAPE-PLD* ($r = 0.81$ and $r = 0.84$, $p_{FDR} < 2.24 \times 10^{-5}$ and $p_{FDR} < 5.4 \times 10^{-6}$, respectively, Fig. 2d).

Functional characterization of blood transcriptomic changes upon diet energy change

Among the 1334 DEG detected in the blood in response to the dietary change, 719 and 615 genes were over- and under-expressed, respectively, in the LE compared to the CT group. KEGG characterization of the over- and under-expressed DEG lists reveals 2 and 8 significantly enriched pathways, respectively ($p_{FDR} < 0.05$) (Additional file 5). The terms for both DEG lists are presented in Table 3.



Gene	cor	Gene	cor	Gene	cor
ENSGALG00000037109	0.892	FADS2	0.838	SDC4	0.821
GLUD1	0.884	DCAF17	0.836	PRTFDC1	0.818
TANK	0.867	ADD3	0.832	USH1C	0.816
BIRC2	0.865	SNX6	0.831	SMU1	0.815
RFTN2	0.859	BDH2	0.830	TRAM1	0.813
DECR1	0.853	LRRC38	0.830	CROT	0.813
SWAP70	0.848	SAR1B	0.828	TBC1D12	0.809
MPHOSPH6	0.846	TUB	0.828	NAPEPLD	0.807
OSTF1	0.843	LOC424740	0.825	CPEB1	-0.806
ABHD3	0.840	DDRGK1	0.821	PLA2G7	0.805

Fig. 2 Lipid metabolism modulation in the hypothalamus in response to the LE diet and genes highly correlated to NR1H3 (LXRα). **a** Schematic summary of the lipid metabolism related genes found to be differentially expressed in the hypothalamus of LE group. **b** Boxplot of the expression of the key lipid transcription factor/nuclear receptors. **c** Top 30 genes which expression is correlated to NR1H3. **d** Co-expression plot of NR1H3 with NAPE-PLD (right) and FADS2 (left). *n.s*: not significant; ***: $p < 0.001$

The pathways associated with under-expressed genes in blood are related to "Metabolic pathways", in particular amino acids biosynthesis (*ACO2*, *ALDH7A1*, *CPS1*, *CTH*, *ENO2*, *GOT1*, *PFKP*, *TALDO1*, *TKT*, *TPI1*), fructose and mannose metabolism (*AKR1B1*, *AKR1B10*, *PFKFB4*, *PFKP*, *PMM2*, *TPI1*) or galactose metabolism (*AKR1B1*, *AKR1B10*, *GALK2*, *PFKP*, *PGM2*). Genes involved in cholesterol biosynthesis were under-expressed in blood (*FDFT1*, *SQLE*, *CYP51A1*, *NSDHL* and *DHCR24*) as in hypothalamus. The two pathways associated with over-expressed genes are "RNA degradation", with *EDC3*, *EXOSC5*, *PABPC1*, *PAN2*, *PAN3*, *PATL1*, *RQCD1*, *SKIV2L* and *TOB2*, and "Ribosome", which contains 3 RPL, 3 MRPL, 3 Ribosomal Protein Lateral Stalk Subunit P (RPLP*x*) and 4 RPS genes, 11 out of these 13 genes were also over-expressed in hypothalamus.

Table 3 KEGG pathways associated with over-expressed (A) and under-expressed DEG (B) in the blood

Term	# of genes	p_{FDR}
A. Under-expressed genes in LE compared to CT		
Metabolic pathways	61	7.92×10^{-05}
Biosynthesis of amino acids	10	2.18×10^{-03}
Carbon metabolism	11	8.02×10^{-03}
Fructose and mannose metabolism	6	9.32×10^{-03}
Steroid biosynthesis	5	9.32×10^{-03}
Amino sugar and nucleotide sugar metabolism	7	9.32×10^{-03}
Pentose phosphate pathway	5	2.20×10^{-02}
Galactose metabolism	5	3.82×10^{-02}
B. Over-expressed genes in LE compared to CT		
Ribosome	13	2.95×10^{-02}
RNA degradation	9	3.24×10^{-02}

Detection of co-expressed genes with WGCNA within hypothalamus and blood DEG lists

To detect gene subsets in our DEG lists, we used the R package WGCNA to identify and cluster co-expressed gene modules (see Methods). As shown in Fig. 3, WGCNA separated for hypothalamus (Fig. 3a) and blood (Fig. 3c) different co-expression groups (noted by a color) for both "LE > CT" (in red) and "LE < CT" (in blue) DEG lists. Interestingly, 2 modules of the same DEG list were not positively correlated in the blood (Fig. 3d, pink color in the correlation matrix) with the blue and purple modules for the red "LE > CT" DEG list and the red and turquoise modules of the blue "LE < CT" DEG list, while all modules were positively correlated in the hypothalamus (Fig. 3b). The plots of module eigengenes of these two pairs can be found in Fig. 3e. We can clearly distinguish in the two plots, two distinct parallel series of points that correspond to the R+ and R- lines. This parallelism reveals two facts: first, a difference of expression between the lines with a positive "R- / R+" expression ratio for the purple module (i.e., the x-axis of the plot in Fig. 3e top) whereas it is negative for the blue module (i.e., the y-axis). Second, the eigengene expression differential between the LE and CT groups (symbolized by a Δ_{diet} in Fig. 3e) is similar for both lines confirming the absence of a diet × line interaction. We found the same characteristics for the red vs. turquoise modules (Fig. 3e bottom). This illustrates again that this difference is independent of the line effect, and the absence of interactions at the gene expression level, as already seen in Fig. 1c.

The functional analysis of each co-expressed gene module in the hypothalamus revealed KEGG terms similar to the full list of over- and under-expressed genes for the turquoise and blue modules, respectively, and no KEGG term enrichment for the green, red and yellow modules. In the pink module, three genes were associated with "N-Glycan biosynthesis", while the brown module was enriched in genes related to vesicles and organelles. Finally, the black module was enriched in terms associated with immunological functions (see Additional file 6). This last module, composed of 134 genes, is associated with 10 immunological-related pathways, supported by 22 genes in total, such as *C1QA*, *C1QB* and *C1QC*, *C3AR1*, *CD14*, *IRF1* and *TLR4*. In the blood, we found seven modules in the list of over-expressed genes and five modules in the list of under-expressed genes. Functional analysis revealed KEGG terms similar to the full list of under-expressed genes for the black module. No KEGG term enrichment were found for the purple, magenta, green, blue, pink, turquoise, brown, and red modules. The greenyellow module was enriched with genes associated to "Ribosome" and "Protein processing in endoplasmic reticulum", while the salmon module was enriched with 3 genes associated with the "Estrogen signaling pathway" (See Additional file 7).

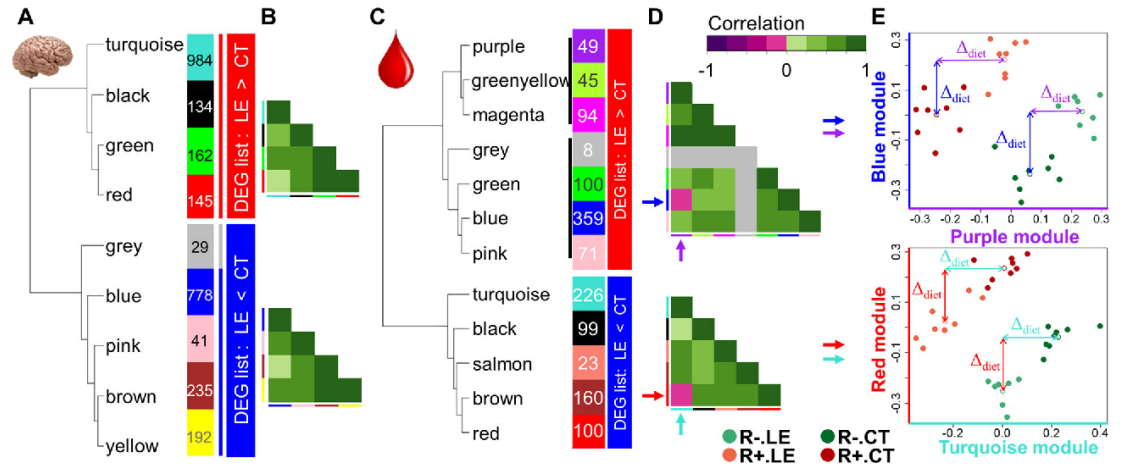


Fig. 3 Analysis of WGCNA modules obtained for the hypothalamus and blood differentially expressed genes. Hierarchical clustering of the eigengenes of the modules detected with hypothalamus (**a**) and blood (**b**) DEG. Module colors are drawn next to module names, with the number of genes in the modules. Unclustered genes are in the grey module. The boxes on the right indicate whether the module contains over-expressed (LE > CT) genes (red) or under-expressed (LE < CT) genes (blue). Black lines highlight the 2 subsets distinguished by WGCNA for the LE > CT DEG list. **c** Heatmap of the correlation matrix between the modules eigengenes. Note the negative correlation (pink boxes) between the purple and blue modules (*top*) and turquoise and red modules (*bottom*). **d** Plots of two pairs of module eigengenes from blood DEG. *Top*: purple vs. blue module from the LE < CT DEG list, *bottom*: turquoise vs. red module from the LE > CT DEG list. Δ_{diet} is the difference between the LE mean vs. CT mean (symbolized with an empty circle) for each line

Focus on genomic regions concentrating differentially expressed genes

We searched for groups of three or more DEG in close physical proximity (i.e., side by side) along the genome that had a significant pairwise expression correlation ($|r| > 0.7$ & $p_{FDR} < 10^{-4}$), hypothesizing that such genes might be co-regulated by a local common mechanism. We found two such proximal co-expressed gene groups in the hypothalamus (Fig. 4a and b), composed of *RPS6KA2*, *MPC1* and *SFT2D1* for the first one (Fig. 4a) and *C1QA*, *C1QB* and *C1QC* for the second (Fig. 4b), genes that belong to the black WGCNA module, which was enriched in immunity-related genes.

Discussion

Layers from both lines adapt to the low-energy diet by increasing feed intake and changing body reserve dynamics

The absence of significant differences in egg production (number and weight) between the LE and CT groups suggests that the animals were able to adapt to a suboptimal diet. The adaptive mechanisms adopted by the animals to compensate for the decrease in diet-energy content involved an increase in feed intake and a decrease of the abdominal adipose tissue. The increase in feed intake in response to a 15%-energy-depleted diet over 14 weeks is consistent with the results from Grobas et al. [1] and Harms et al. [2]. However, this increased ingestion did not allow the layers from the LE group to fully compensate for the difference in energy, (Table 1) as indicated by the significant difference in Energy Intake between the diet groups. The decrease of the percentage of fat weight to the total weight, probably resulting from this incomplete compensation, is consistent with the results reported by Murugesan and Persia [3], where layers were fed a 3%-energy-depleted diet compared to the control over 11 weeks, although the authors did not observe a feed intake modification, perhaps due to the small difference in energy between the two diets.

The absence of a significant line × diet interaction at the expression level is consistent with the absence of interaction at the trait level, meaning that both R+ and R- birds reacted to the energy-depleted diet in a similar way and with the same magnitude. At the expression level, the Δ_{diet} values in Fig. 3e illustrates this conclusion: as an example, Δ_{diet} for the genes belonging to the purple module are similar in the two lines whereas these genes are more expressed in R- than in R+.

Liver and adipose tissue transcriptomes were unaffected by the low-energy diet

Neither the abdominal adipose tissue nor the liver transcriptomes were affected by the diet change, as shown by the small number of differentially expressed genes in these two tissues (15 and 2, respectively). The absence of differentially expressed genes in the abdominal adipose tissue indicates that the mobilization of body reserves observed with the adipose tissue weight decrease was not mainly driven transcriptionally. This observation is consistent with the fact that the two key genes of adipocytes lipolysis, *PNPLA2* (*alias ATGL*) coding the enzyme catalyzing the initial step of this process and *LIPE* coding the Hormone-Sensitive Lipase which primarily hydrolyzes stored triglycerides to free fatty acids are known to be quickly regulated through post-translational modifications such as phosphorylation [6]. We further confirmed that these two genes were not differentially expressed using RT-qPCR (for *PNPLA2*, $\Delta Ct_{LE\text{-}CT} = 0.02$,

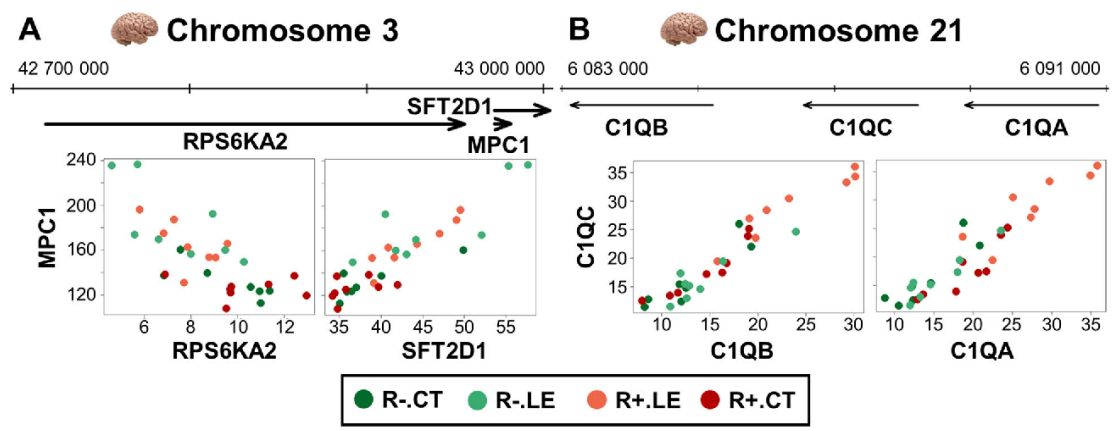


Fig. 4 Genomic localization and pairwise scatterplots of expression of four groups of co-expressed and co-localized genes. In each plot, *top:* genomic localization of the three genes. *Bottom*: pairwise scatterplots of expression (FPKM) the genes. **a** cluster composed of *RPS6KA2*, *MPC1* and *SFT2D1*. **b** cluster composed of *C1QA*, *C1QB* and *C1QC*

$p = 0.97$ and for *LIPE*, $\Delta Ct_{LE\text{-}CT} = 0.21$, $p = 0.50$). The mobilized lipids resulting of this probable adipose tissue lipolysis could have been used by the hypothalamus as an energy source, as we discuss later. Concerning the liver, the absence of reaction at the transcriptomic level shows that the difference in energy between the two diets did not impact gene expression, which suggests an absence of hepatic lipid metabolism variation. Indeed, lipid metabolism is known to be highly regulated at the transcriptional level, as previously shown in chickens [5, 7]. In these studies, which explored the impact of the diet fiber and lipid composition variation or the fasting-feeding transition (known to impact hepatic lipid metabolism), numerous genes involved in the lipid metabolism were impacted at the transcriptional level. The result observed here in liver can be explained by the partial compensation of the energy depletion by the increase in feed intake and the mobilization of the body reserves. We confirmed by RT-qPCR the absence of differential expression of *PPARα*, a key genes of fatty acid *β*-oxidation ($\Delta Ct_{LE\text{-}CT} = -0.16$, $p = 0.30$) and for *FASN* and *SREBF1*, two key genes of fatty acid synthesis (for *FASN*, $\Delta Ct_{LE\text{-}CT} = -0.24$, $p = 0.37$ and for SREBF1, $\Delta Ct_{LE\text{-}CT} = -0.14$, $p = 0.57$).

Blood cells participate in the adaptation to the CT versus LE diet changes

While the liver and adipose tissue were almost unaffected by the low energy diet, at least from the transcriptomic point of view, blood cell genes reacted strongly to the low-energy diet with more than 1000 genes modulated by the diet change but for which the interpretation remains difficult. Indeed, the red blood cell components differ between mammals and vertebrates more distant in the evolutionary scale, such as birds or fish. In these animals, erythrocytes and thrombocytes are nucleated and their transcriptional activity is not yet well defined. Secondly, the blood transcriptome is mainly studied to evaluate the response to an inflammatory and immune challenge and rarely to study the effects of diets. To our knowledge, no study has explored so far the blood transcriptome profile in chicken under such conditions. We found an activation of genes involved in RNA degradation and ribosome activity and a repression of genes involved in cholesterol and amino acid biosynthesis, as well as galactose and fructose metabolisms. Cholesterol synthesis decrease in response to energy restriction was also reported by Bouvier-Muller et al. [8] in energy-restricted ewe's blood transcriptome. Under-expression of some of the genes described in our study like *CYP51A1*, *DHCR24*, *FDFT1* and *SQLE* was also observed in ewes fed a low energy diet versus control (restriction to 60% of the calculated net energy requirements during 15 days). Furthermore, three genes involved in macrophage cholesterol efflux and transport [9] show a significant, or a trend toward, over-expression in our study: *ABCA1* (FC = 1.68 $p_{FDR} = 0.07$) and *APOA1* (FC = 2.10 $p_{FDR} = 0.08$), the latter being the chicken equivalent of human *APOE* [10], and *CETP* (FC = 1.61, $p_{FDR} = 0.02$). The precise relationship between these genes and their differential expression remains to be linked with the feed intake. Taken together, these reports and our results suggest that the chicken blood transcriptome may play a role in the adaptation of birds to feed stress. However, the differentially expressed genes are quite hard to interpret, and further studies will be required to unveil the mechanisms at play.

In the hypothalamus, the low-energy diet seems to alter the general synaptic organization, partly through a modulation of cholesterol and a global protein synthesis associated to fatty acid *β*-oxidation

The hypothalamus is a brain area that integrates metabolic and hormonal cues and controls appetite and peripheral metabolism. It is composed of different cell types, including neurons and "non-neuronal" cells (such as astrocytes, microglial cells, oligodendrocytes and endothelial cells) [11], and the transcriptomic changes observed in this study reflect most likely changes occurring in different cells, but we are unable to distinguish which ones. Notwithstanding, the differential expression analysis suggests an effect of the low-energy diet in neuronal circuits. We detected an under-expression of genes involved in the synaptic vesicle cycle, as well as in the glutamatergic, dopaminergic and GABAergic synapses. In addition, key genes involved in the cholesterol synthesis (*CYP51A1*, *DHCR7*, *DHCR24*, *FDFT1* and *SQLE*) and in the cholesterol efflux from neuronal cells, namely *ABCA7* (FC = 0.67, $p_{FDR} = 0.03$) and *ABCG4* (FC = 0.64, $p_{FDR} = 0.007$) [12] were also under-expressed. Interestingly, the adult brain is the most cholesterol-rich organ, containing 20% of the whole body's cholesterol [13]. The majority of it is present in myelin sheaths and the rest in the plasma membranes of astrocytes and neurons to maintain their morphology and synaptic transmission [14]. Taken together, these findings reveal a link between nutrition and brain plasticity in chicken, as it has already been described in mice [15, 16]. Furthermore, our results suggest an overall activation of protein synthesis in the hypothalamic cells, one of the most energy-consuming processes in a cell [17], probably reflecting the protein machinery necessary to promote feed intake increase. Indeed, we detected in the hypothalamus of the low-energy group 83 over-expressed DEG related to the ribosome machinery indicating activation of numerous genes related to the oxidative phosphorylation (that produces ATP) and the fatty acid oxidation (used as fuel for the respiratory chain) (Fig. 5). Concerning the oxidative phosphorylation, we observed 32 over-expressed genes coding the 5 protein complexes located in the inner

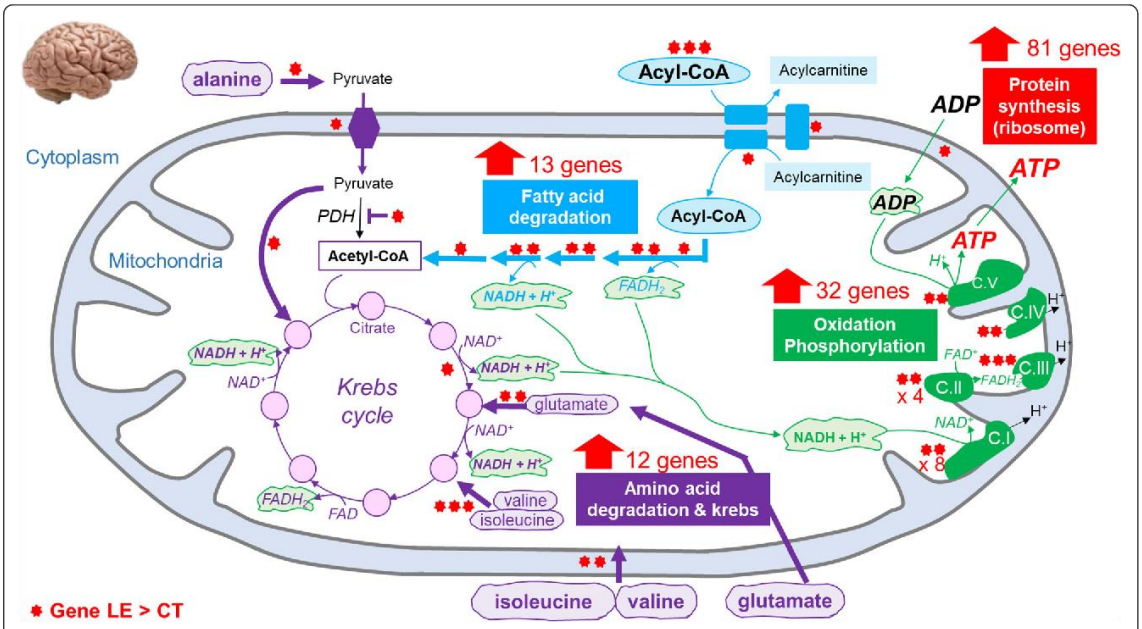


Fig. 5 Proposed mechanism of energy pathways increased in the hypothalamic cells in LE diet. In blue: reactions related to fatty acid β-oxidation (*ETFDH*, *ACADL*, *ACADS*, *ECHS1/ECI1*, *HADH*, *HADHB*, *ACAA2*), to fatty acids transport through the plasma (*FABP4*, *FABP7*, *SLC27A1*), and the mitochondrial (*CPT2*, *CACT*) membrane. In purple: reactions related to TCA cycle (*IDH2*), to transport of amino-acids (*BCKDHA*, *BCKDHB*) and pyruvate (*MPC1*, *MPC2*) in the mitochondria, to the integration of amino-acids in the TCA cycle as α-ketoglutarate (*GDH1*, *GPT2*) or succinyl-CoA (*ALDH6A1*, *ECHS1*, *HIBDCH*) and of the pyruvate as oxaloacetate (*PC*). In green: reactions related to oxidative phosphorylation and mitochondrial respiratory chain complex I (*MT-ND1*, *MT-ND2*, *MT-ND3*, *MT-ND4*, *ACAD9*, *MT-ND4L*, *MT-ND5*, *MT-ND6*, *NDUFA2*, *NDUFA8*, *NDUFA10*, *NDUFB9*, *NDUFS4*, *NDUFV3*, *FOXRED1*), complex II (*MT-CO1*, *MT-CO2*, *MT-CO3*, *APOPT1*, *COX14*, *COX7B*, *COA5*, *COA6*), complex III (*MT-CYB*, *UQCRB*, *UQCRQ*), complex IV (*MT-ATP6*, *MT-ATP8*) and complex V (*SDHD*, *SDHAF2*), as well as the entry of ADP in the mitochondria (*SLC25A1*)

mitochondrial membrane (Fig. 5) including the ADP/ATP translocase 1 (*SLC25A4*, FC = 1.79, $p_{FDR} = 3.31 \times 10^{-06}$) required for the entry of ADP (the substrate of the ATPase) in the mitochondria, and considered as a limiting factor of this process. The NADH and $FADH_2$ required by the respiratory chain is produced by the mitochondria β-oxidation of fatty acid, which increase is supported by 10 over-expressed DEG (Fig. 5) and by the integration of amino acids in the Krebs cycle as indicated by the 12 over-expressed DEG identified (Fig. 5). While short and medium chain fatty acids appear to enter the brain-blood barrier by simple diffusion through the plasma membrane, long chain fatty acids (> 12 carbons) need transporters to cross the brain-blood barrier. Some of these transporters such as *FABP4, FABP7* [18] and *SLC27A1* [19] were also overexpressed. Cedernaes et al. [20], obtained similar results, although in a different context. The authors observed an over-representation of genes related to oxidative phosphorylation as well as to ribosome sub-units in mice hypothalamus following a fasting period, and others studies [21] made a link between mitochondrial oxidation of fatty acids in the hypothalamus and increase in feed intake.

Hypothalamic arachidonic acid may be involved in the difference of feed intake between LE and CT groups through mechanisms involving the hypothalamic endocannabinoid and complement systems

The involvement of the endocannabinoids in the regulation of feed intake is well documented [22, 23], in particular for the two best known representative of this family of molecules, the 2-AG (2-arachidonoylglycerol) and the Arachidonoyl ethanolamine (AEA also called Anandamine). Both these molecules are the ligands of the endocannabinoid receptor, CN1R. Interestingly, we observed an under-expression of *DAGLB* (FC = 0.74, $p_{FDR} = 0.003$), involved in the synthesis of 2-AG [24] and an over-expression of *MGLL* (FC = 1.75, $p_{FDR} = 5.73 \times 10^{-06}$), coding an enzyme responsible for 2-AG degradation. We also observed an over-expression of *NAPE-PLD* (FC = 1.95, $p_{FDR} = 6.86 \times 10^{-11}$), which codes for the enzyme that catalyzes the second step of the classical "two-step" pathway of the synthesis of AEA and other NAEs. The first step of this pathway consists in the formation of N-acylphosphatidyl ethanolamines (NAPEs) by the transfer of the acyl chain of phospholipids on phosphatidylethanolamine by a calcium-dependant transacylase [25]. *NAPE-PLD* then catalyses the

cleavage of NAPEs to yield NAEs. Different NAEs are generated depending on the nature of the acyl chain in the first step. For example, Arachidonoyl ethanolamine (AEA) derives from the ω6 poly-unsaturated fatty acid (PUFA) arachidonate and Palmitoyl ethanolamide (PEA) derives from the saturated fatty acid palmitate [26]. We observed an over-expression of *FADS1* (FC = 1.99, $p_{FDR} = 3.25 \times 10^{-14}$), *FADS2* (FC = 2.07, $p_{FDR} = 3.15 \times 10^{-10}$), *ELOVL2* (FC = 1.87, $p_{FDR} = 0.003$) and *ELOVL5* (FC = 1.48, $p_{FDR} = 0.0004$), key genes of the PUFA ω6 synthesis [27]. FADS2 catalyzes the Δ6-desaturation of the essential fatty acid linoleic acid (C18:2 ω6) into γ-linolenic acid (C18:3ω6), which is elongated into C20:3 ω6 by ELOVL5; the C20:3 ω6 is then Δ5-desaturated into arachidonic acid (C20:4 ω6) by FADS1 [28] which may lead to the formation of AEA [29]. As 2-AG, AEA could also activates the CB1R endocannabinoid receptor, leading to an increase of feed intake [30]. Consistently with this hypothesis, we observed an under-expression of CB1R which might be due to a negative feedback following CB1R activation. Figure 6 summarizes this proposed mechanism. Interestingly, we found that *FADS2* and *NAPE-PLD* were highly correlated to *NR1H3* (alias *LXRα*) that codes for a receptor involved in the control of various physiological functions with a major role in fatty acid homeostasis and cholesterol metabolism [31]. The mechanism of the regulation of the *FADS2* and *NAPE-PLD* transcription that can be direct or indirect, remains to be elucidated. Interestingly, the arachidonic acid is also a precursor of the prostaglandins [32], which has been shown to be involved in feed intake regulation along with complement system molecules [33].

Among the eight modules detected by WGCNA using the lists of hypothalamic DEG, the black module was composed of over-expressed genes related to immunity. Three of them, *C1QA*, *C1QB* and *C1QC* were detected as co-localized and co-expressed genes. The co-localization and strong co-expression of *C1QA*, *C1QB* and *C1QC* strongly suggest a mechanism of common regulation. These three genes code for the A, B and C polypeptide chains composing the C1q molecule, a subcomponent of the C1 complex involved in the complement activation [34]. The complement system is a part of the innate immune system, involved in the host defense against bacteria and in the removal of wastes [35]. C3AR1, the receptor of C3a, which is produced upon the activation of the complement system [36], also belongs to the black co-expression module. Interestingly, Ohinata et al. showed that an agonist of C3AR could suppress feed intake in mice [37] through prostaglandin (PG) E_2 production [33]. Furthermore, the same authors showed that C5a, another member of the complement system, stimulated feed intake via a mechanism involving this time PGD_2 [38]. Interestingly, as we

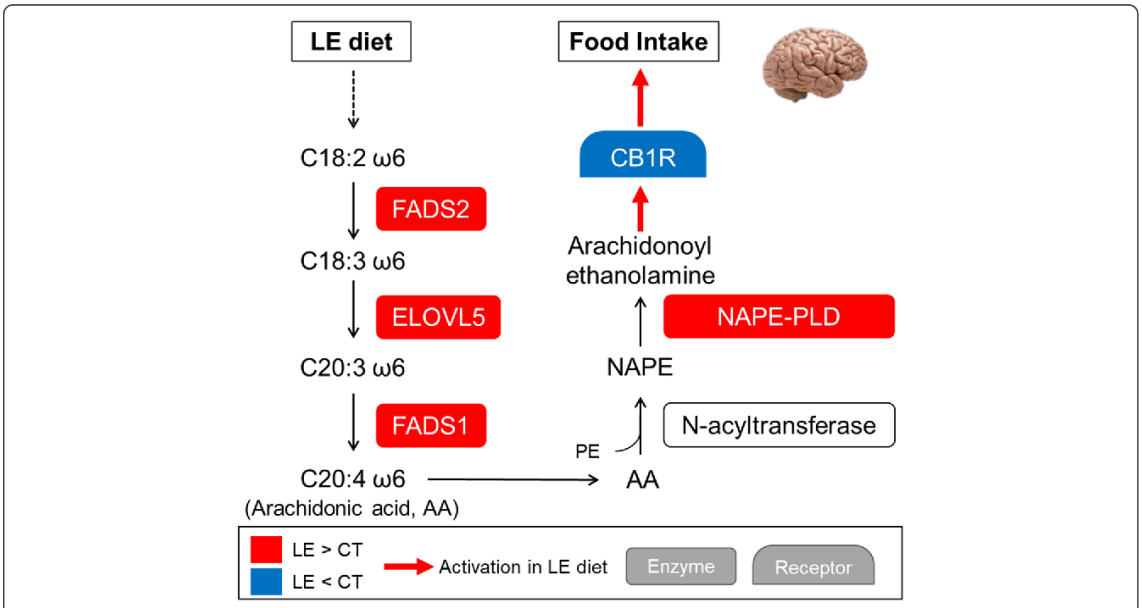


Fig. 6 Proposed mechanism leading to an increased feed intake in the LE diet. Diet fatty acids are processed by *FADS1*, *FADS2*, *ELOVL5* and *FADS1*, leading to the production of arachidonic acid (AA). The arachidonic acid eventually lead to the production of Arachidonoyl ethanolamine (AEA), thanks to the action of NAPE-PLD. The AEA acts on CB1R, leading to an increase in feed intake. FADS1 and 2: Fatty Acid Desaturase 1 and 2, ELOVL5: Elongation Of Very Long Chain Fatty Acids Protein 5, NAPE-PLD: N-Acyl Phosphatidylethanolamine Phospholipase D, CB1R: Cannabinoid Receptor 1, AA: Arachidonic Acid, PE: Phosphatidylethanolamine, NAPE: N-arachidonoyl phosphatidylethanolamine, AEA: Arachidonoyl ethanolamine (alias Anandamide)

discussed earlier, we found that key genes of the poly-unsaturated fatty acid (PUFA) ω6 synthesis, that lead to the formation of arachidonic acid, the precursor of prostaglandins, were overexpressed in LE group. Finally, *C1QTNF4* (C1q/TNF-related Protein 4), that possesses two tandem globular C1q domains and is under-expressed in LE versus CT (FC = 0.65, $p_{FDR} = 0.01$), has also been shown to suppress feed intake in mice [39]. Surprisingly, we found only one other group of 3 co-localized and co-expressed genes in the hypothalamus DEG lists. Such results show that regulatory mechanisms affecting different genes located in a same genomic region are not so frequent in response to a diet change despite the high number of DEG identified and analyzed in the hypothalamus and the blood. We found similar results in a previous study that evaluated on the impact of diet-composition change on the breast muscle, adipose tissue and liver of broiler, in which one region was identified [5].

Conclusions

This work is the first to provide a multi-tissue analysis of layers submitted to a hypo-energetic diet on a long period. Neither the adipose tissue nor the liver seemed to be affected by the diet change at the transcriptional level, suggesting regulations occurring at a different level. In contrast, we observed a strong effect of the diet on the hypothalamic transcriptome of the layers. The regulation of feed intake in the hypothalamus is a complex mechanism. Our results in chicken suggest, as in mice, a link between feed intake and brain plasticity, as well as fatty acid metabolism [40–43]. We show here a mechanism in chickens that seems to modify feeding behavior through an increase in feed intake in response to a low-energy diet, allowing egg mass production to be maintained, probably through the action of the endocannabinoid and the complement systems that involve the hypothalamic poly-unsaturated fatty acid synthesis, and in particular the arachidonic acid. Overall, this work contributes to a better understanding of the adaptive strategies employed by chickens to cope with a suboptimal diet and the impact that this suboptimal feeding may have on egg quality and production. Such understanding is of importance in the frame of the globalized poultry market, in which commercial animals are exposed to a wide diversity of production conditions.

Methods

Animals and diet

Laying hens were hatched at the INRA Pôle d'Expérimentation Avicole de Tours (PEAT) in Nouzilly, France. They belonged to two Rhode Island Red layer lines that underwent a 40-year diverging selection on residual feed intake (RFI) [4]. The RFI represents the difference between the observed and the predicted feed consumption based on a multiple regression equation taking into account the average body weight, the weight variation and, for females, the mass of eggs produced over a given period [44, 45]. The R+ chickens were selected to have a positive RFI, reflecting a low feed efficiency, while the R- chickens were selected to have a negative RFI and therefore to be feed efficient. They were reared under standard farming conditions in floor pens until 17 weeks of age. At this age, 45 R+ and 51 R- hens were transferred in individual cages and reared under thermo-neutral conditions (22 °C), with a lighting regimen set at 14 h of light per day and an ad libitum feeding. Of these, 34 R+ and 36 R- hens were fed a commercial diet (control group, CT) and 11 R+ and 15 R- were fed a low-energy diet (low-energy group, LE). The two diets had a similar protein content, while the energy content was reduced by 15% in the LE diet as compared to the standard diet (2450 kcal/kg versus 2880 kcal/kg), due to the replacement of soybean and maize by rapeseed and raw wheat, and by increasing the raw cellulose percentage (7.4 g/kg against 2.6 g/kg). The composition of both diets is detailed in Additional file 8.

Tissue sampling

At 31 weeks, eight animals from each line (R- and R+) and from each diet (CT and LE) were selected as representative of the group for slaughtering, that is 8 × 2 × 2 = 32 animals. Layers were slaughtered at the fed status by neck cut and bleeding, immediately after head electrical stunning. Right after slaughter, abdominal adipose tissue, the extremity of the left liver lobe and hypothalamus were sampled, snap frozen in liquid nitrogen and stored at − 80 °C until analysis. Blood samples from the same animals were collected from the occipital sinus in EDTA tubes and 100 μL of blood were removed and diluted in 1 mL of TRIzol® reagent (Invitrogen, California, USA). After a vigorous agitation, the tube was maintained at room temperature for five minutes, then quickly frozen in liquid nitrogen and stored at − 80 °C until RNA extraction.

Traits collection and analysis

Seven traits related to performance and body composition were recorded for the 45 R+ (34 CT and 11 LE) and 51 R- (36 CT and 15 LE) birds. Egg number was recorded from the date of the first egg (around 21 weeks of age) to 31 weeks of age and laying rate (i.e. number of egg laid during the recording period divided by the length of the period in day, expressed in %) was calculated; egg weight (g), static stiffness ($N.mm^{-1}$) were calculated from 3 eggs per hen collected at 30 weeks of age, and abdominal adipose was weighted at slaughter. Weekly feed intake was measured over 4 weeks, from 27 to 31 weeks of age and body weight (g) at 31 weeks of age. Residual feed intake was computed as described in Bordas et al. [4]. Traits were

analyzed with R version 3.4.2 [46]. A two-way analysis of variance was performed with line, diet and the interaction between line and diet as main effects using the R function lm, and the R package "car" [47].

RNA isolation

Approximately 100 mg of adipose tissue and 30 mg of liver were homogenized in TRIzol® reagent (Invitrogen, California, USA), and the whole blood mixed with 1 mL of TRIzol® was adjusted between 4 and 4.5 with 10 μL of 5 N glacial acetic acid [48]. The total RNA was then extracted according to the manufacturer's instructions, resuspended in 50 μL of RNA-free water and stored at − 80 °C. For the hypothalamus, we used the kit Allprep DNA/RNA (Qiagen). The RNA was extracted from the hypothalamus according to the manufacturer's instructions. The total RNA was quantified with a NanoDrop® ND-1000 spectrophotometer (Thermo Scientific, Illkirch, France). The RNA quality was controlled using an Agilent 2100 bioanalyzer (Agilent Technologies France, Massy, France). The average RNA integrity numbers were 7.3 ± 0.6 (mean ± SD) for the adipose tissue, 8.8 ± 0.48 for the hypothalamus, 8.2 ± 0.5 for the whole blood and 9.2 ± 0.3 for the liver.

RNA-seq data acquisition

Paired-end sequencing was conducted on all samples using an Illumina HiSeq3000 (Illumina, California, USA) system, with 2 × 150 bp. Libraries with an on average 465-bp insert were prepared following Illumina's instructions by purifying poly-A RNAs (TruSeq RNA Sample Prep Kit). Illumina adapters containing indexing tags were added for subsequent identification of samples. Samples were PCR-amplified, and quantitative PCR was then performed for library quantification (QPCR NGS Library Quantification kit). Eight samples were filled on one lane within a flow cell with 2 samples for each of the four line × diet groups to minimize the inter-lane bias. After sequencing, the indexed adapter sequences were trimmed using CASAVA v.1.8.2 software (Illumina). We obtained an average of 90 million reads per sample (84 million for the adipose tissue, 98 million for the blood, 86 million for the hypothalamus and 90 million for the liver), for a grand total of 11 billion reads. For each sample, reads were mapped to the *Gallus gallus*-5 reference genome using STAR v.2.3.0e [49]. PCR duplicates were removed using rmdup tool from SAMtools suite [50]. For each sample, quantification was performed using RSEM [51] with the Ensembl v93 annotation.

RNA-seq data analysis

All the analyses were performed with R version 3.4.2. The trimmed mean of M-values (TMM) scaling factor method was used for library size normalization [52] using the R/Bioconductor package edgeR [53] version 3.12.1. In each tissue, the expressed genes were selected if their FPKM expressions were over 0.1 in at least 80% of the samples of a group line × diet (FPKM expression being obtained after TMM normalization using "rpkm" function from edgeR package). Differential expression analysis was performed using the R/Bioconductor package edgeR [53] based on a generalized negative binomial model for model fitting. We used the "edgeR-Robust" method to account for potential outliers when estimating per gene dispersion parameters [54]. *P*-values were corrected for multiple testing using the Benjamini-Hochberg approach [55] to control the false discovery rate (FDR), and genes were identified as significantly differentially expressed if $p_{FDR} < 0.05$.

Functional enrichment analysis

The enrichment analysis of Kyoto Encyclopedia of Genes and Genomes (KEGG) terms in each list of interest of differentially expressed genes was performed using the STRING tool [56] (https://string-db.org). Only the 1-to-1 human orthologous genes with a standardized HGNC name were submitted for the analysis, i.e. 67.4% of the 18,346 protein-coding genes of chicken Ensembl v93 annotation.

Co-expression module detection with WGCNA

We used the R package WGCNA [57] to detect co-expression modules based on gene expression data and a weighted correlation network. Briefly, WGNCA screens for clusters (called modules) of highly correlated genes in the expression dataset. Indeed, while within a list of over- or under-expression in one condition versus another one, one can expect all the genes to be positively correlated to one another, such list can be split into modules of genes with a higher expression correlation among them than with the rest of the list. These genes are more likely to share a common regulation and a common biological function and therefore may highlight more specifically one pathway. In addition, it may happen that a gene subset is not correlated with the other subsets of the same DEG list because of factors other than the one used for the differential expression analysis. These modules are summarized by an eigengene, which corresponds to the first principal component of the module. These eigengenes enable comparisons between modules, clustering of modules, and calculations of correlations between modules and phenotypes. Modules hierarchical clustering was realized using as "1 − the pearson correlation" between modules as distance criterion and "ward's" method as aggregation criterion.

Detection of co-localized differentially expressed genes

We used R home-made script to screen for groups of three or more differentially expressed genes, located side-by-side, without consideration for distance, and with a significant pairwise Spearman expression correlation ($|r| > 0.7$ and $p_{FDR} < 10^{-4}$).

RT-qPCR analysis

Reverse transcription (RT) was carried out using the high-capacity cDNA archive kit (Applied Biosystems, Foster City, CA) according to the manufacturer's protocol. Briefly, reaction mixture containing 2 μL of 10× RT buffer, 0,8 μL of 25X dNTPs, 2 μL of 10X random primers, 1 μL of MultiScribe Reverse Transcriptase (50 U/ μL), and total RNA (2 μg) was incubated for 10 min at 25 °C followed by 2 h at 37 °C and 5 min at 85 °C. Dilution RT reaction was further used for real time quantitative PCR (qPCR). 5 μl of cDNA samples were mixed with 7,5 μl of Sso Advanced Universal SYBR Green Supermix (Bio-Rad), 1,5 μl H20 and 330 nM of specific reverse and forward primers. Reaction mixtures were incubated in an CFX connect Real-Time PCR Detection System (Bio-Rad, Marne la Coquette, France) programmed to conduct one cycle (95 °C for 30 s) and 43 cycles (95 °C for 15 s and 60 °C for 30 s). A melting curve program was then performed for each gene to check the presence of a unique product with specific melting temperature. For each sample and each gene, PCR runs were performed in duplicates. The sequences of the primers used were, from 5′ to 3′: *LIPE*, forward "GTCTCGGGTTCCAGTTCGTG", reverse "CGTAGGACACCAACCCGATG". *PNPLA2*, forward "TGGGCAGTCATCTTTCAGCCA", reverse "AAGCTGACGCTGGTACTCCT". *FASN*, forward "TGAAGGACCTTATCGCATTGC", reverse "GCATGGGAAGCATTTTGTTGT". *PPARα*, forward "GTCGCTGCCATCATTTGCTGT", reverse "TTGCCGGAGGTCAGCCATTT". SREBF1, forward "GTCGGCGATCCTGAGGAA", reverse "CTCTTCTGCACGGCCATCTT".

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-019-6384-8.

Additional file 1. Means (±SD) and significance for Body weight at 17 weeks for the effect of the diet, the line and their interaction.

Additional file 2. List of the genes uniquely expressed in each tissue.

Additional file 3. List of the differentially expressed genes (LE vs. CT) in each tissue.

Additional file 4. List of the 26 and 44 KEGG pathways significantly enriched in the over- and under-expressed genes in the hypothalamus.

Additional file 5. List of the 2 and 8 KEGG pathways significantly enriched in the over- and under-expressed genes in the blood.

Additional file 6. List of the KEGG pathways significantly associated with WGCNA's modules detected using hypothalamic DEG.

Additional file 7. List of the KEGG pathways significantly associated with WGCNA's modules detected using blood DEG.

Additional file 8. Composition of the diets.

Abbreviations

AA: Arachidonic Acid; ACAA1: Acetyl-CoA Acyltransferase 1; ACADL: Acyl-CoA Dehydrogenase Long Chain; ACADS: Acyl-CoA Dehydrogenase Short Chain; ACO2: Aconitase 2; ACOX2: Acyl-CoA Oxidase 2; ACSBG1: Acyl-CoA Synthetase Bubblegum Family Member 1; AEA: Arachidonoyl ethanolamine (alias Anandamide); ALDH7A1: Aldehyde Dehydrogenase 7 Family Member A1; APOA1 and APOC3: Apolipoproteins; ATP5G1: ATP Synthase Membrane Subunit C Locus 1; ATP5H: ATP Synthase Peripheral Stalk Subunit D; ATP5J: ATP Synthase Peripheral Stalk Subunit F6; C1QA, C1QB and C1QC: Complement C1q chains; C3AR1: Complement C3a Receptor 1; CB1R: Cannabinoid Receptor 1; CD14: CD14 Molecule; CPS1: Carbamoyl-Phosphate Synthase 1; CPT2: Carnitine Palmitoyltransferase 2; CT: Control; CTH: Cystathionine Gamma-Lyase; CYP51A1: Cytochrome P450 Family 51 Subfamily A Member 1; DBI (alias ACBP): Diazepam Binding Inhibitor (alias Acyl-CoA Binding Protein); DEG: Differentially Expressed Genes; DHCR24: 24-Dehydrocholesterol Reductase; EDC3: Enhancer Of MRNA Decapping 3; ELOVL2 and ELOVL5: Elongation Of Very Long Chain Fatty Acids Protein; ENO2: Enolase 2; EXOSC5: Exosome Component 5; FABP7: Fatty Acid Binding Protein 7; FADS1 and FADS2: Fatty Acid Desaturases; FDFT1: Farnesyl-Diphosphate Farnesyltransferase 1; FDR: False Discovery Rate; FPKM: Fragment Per Kilobase Million; GABA: gamma-Aminobutyric acid; GOT1: Glutamic-Oxaloacetic Transaminase 1; HIF1: hypoxia-inducible factor-1; IRF1: Interferon Regulatory Factor 1; KEGG: Kyoto Encyclopedia of Genes and Genomes; LE: Low-Energy; ME1: Malic Enzyme 1; MRPL: Mitochondrial Ribosomal Protein L; MRPS: Mitochondrial Ribosomal Protein S; MT-CO1 to MT-CO3: Mitochondrially Encoded Cytochrome C Oxidases; MT-CYB: Mitochondrially Encoded Cytochrome B; MT-ND1 to MT-ND6: Mitochondrially Encoded NADH:Ubiquinone Oxidoreductase Core Subunits; NAE: N-acyl ethanolamine; NAPE: N-arachidonoyl phosphatidylethanolamine; NAPE-PLD: N-Acyl Phosphatidylethanolamine Phospholipase D; NR1H3 (alias LXRα): Nuclear Receptor Subfamily 1 Group H Member 3 (alias Liver X Nuclear Receptor alpha); NSDHL: NAD(P) Dependent Steroid Dehydrogenase-Like; PABPC1: Poly(A) Binding Protein Cytoplasmic 1; PAN2 and PAN3: Poly(A) Specific Ribonuclease Subunit PANs; PE: Phosphatidylethanolamine; PEA: Palmitoylethanolamide; PFKP: Phosphofructokinase, Platelet; PLPP5: Phospholipid Phosphatase 5; RFI: Residual Feed Intake; RPL: Ribosomal Protein L; RPS: Ribosomal Protein S; RQCD1: CCR4-NOT Transcription Complex Subunit 9; SCP2: Sterol Carrier Protein 2; SDHD: Succinate Dehydrogenase Complex Subunit D; SKIV2L: Ski2 Like RNA Helicase; SLC27A1: Solute Carrier Family 27 Member 1; SQLE: Squalene Epoxidase; TALDO1: Transaldolase 1; TKT: Transketolase; TLR4: Toll Like Receptor 4; TMM: Trimmed Mean of M-values; TOB2: Transducer Of ERBB2, 2; TPI1: Triosephosphate Isomerase 1; WGCNA: Weighted Gene Co-expression Network Analysis; WHSC1L1 (alias NSD3) : Nuclear Receptor Binding SET Domain Protein 3

Acknowledgements

The authors thank the staff of the INRA experimental poultry unit (UE1295 PEAT, Nouzilly, France) for producing and rearing animals, and the technicians of the research units for helping to measure birds. The authors also thank F. Boissel and M. Bellier for helping with RNA extraction, M. Lessire for valuable advices in defining the LE diet.

Authors' contributions

FP, TZ, AC and SLagarrigue conceived the experimental design. TZ and SLagarrigue coordinated the study. FJ, CD, MB, SLeroux, KM, DG, FP, AC, TZ, and SLagarrigue participated to the set-up of the experimental design and sample collection; MBoutin, CD, and SLeroux carried out all RNA extractions. TB defined the diet. DE generated RNA-seq libraries and sequencing; CK performed bioinformatics pre-processing of the RNA-seq data. FJ carried out all RNA-seq analyses, with assistance from MBrenet. FJ, AR, YB, TZ and SLagarrigue conceived and/or participated in the statistical analyses; FJ, AB, TZ and SLagarrigue interpreted the data; LL realized and interpreted RT-qPCR analysis added in the revised version. FJ, TZ and SLagarrigue drafted the manuscript. All authors helped to draft the manuscript and read and approved the final version.

Funding

This project received financial support from French National Agency of Research (ChickStress Project, ANR-13-ADAP) and from the European Union's H2020 program under grand agreement no. 633531 (Feed-a-Gene project). FJ and KM are Ph.D. fellows supported by the Brittany region (France) and the INRA Animal Genetics division. These funding bodies had no role in the design of the study, in the collection, analysis, and interpretation of data, or in writing the manuscript.

Availability of data and materials
The 64 RNA-seq samples are available in European Nucleotide Archive (ENA) through ENA Series accession number PRJEB28745.

Ethics approval and consent to participate
The experiments was conducted at the experimental fame PEAT under license number C37–175-1 for animal experimentation, in compliance with the European Union Legislation, and was approved by the local ethical committee in animal experimentation (Val de Loire) and by the French Ministries of Higher Education and Scientific Research, and of Agriculture and Fisheries (n°2873–2015112512076871).

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]PEGASE UMR 1348, INRA, AGROCAMPUS OUEST, 35590 Saint-Gilles, France. [2]SIGENAE Plateform, INRA, 31326 Castanet-Tolosan, France. [3]GABI UMR 1313, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France. [4]GenPhySE UMR 1388, INRA, INPT, ENVT, Université de Toulouse, 31326 Castanet-Tolosan, France. [5]Programme Cartes d'Identité des Tumeurs (CIT), Ligue Nationale Contre Le Cancer, 75013 Paris, France. [6]GENOTOUL Plateform, INRA, 31326 Castanet-Tolosan, France. [7]PEAT UE, INRA, 37380 Nouzilly, France. [8]NOVOGEN, Mauguérand, 22800 Le Foeil, France. [9]BOA UMR, INRA, Université de Tours, 37380 Nouzilly, France. [10]Centre des Sciences du Goût et de l'Alimentation, AgroSup Dijon, CNRS, INRA, Université de Bourgogne, Dijon, France.

Received: 13 May 2019 Accepted: 11 December 2019
Published online: 30 December 2019

References

1. Grobas S, Mendez J, De Blas C, Mateos G. Laying hen productivity as affected by energy, supplemental fat, and linoleic acid concentration of the diet. Poult Sci. 1999;78:1542–51.
2. Harms RH, Russell GB, Sloan DR. Performance of four strains of commercial layers with major changes in dietary energy. J Appl Poult Res. 2000;9:535–41.
3. Murugesan GR, Persia ME. Validation of the effects of small differences in dietary metabolizable energy and feed restriction in first-cycle laying hens. Poult Sci. 2013;92:1238–43.
4. Bordas A, Tixier-Boichard M, Merat P. Direct and correlated responses to divergent selection for residual food intake in Rhode island red laying hens. Br Poult Sci. 1992;33:741–54.
5. Desert C, Baéza E, Aite M, Boutin M, Le Cam A, Montfort J, et al. Multi-tissue transcriptomic study reveals the main role of liver in the chicken adaptive response to a switch in dietary energy source through the transcriptional regulation of lipogenesis. BMC Genomics. 2018;19. https://doi.org/10.1186/s12864-018-4520-5.
6. Kim S-J, Tang T, Abbott M, Viscarra JA, Wang Y, Sul HS. AMPK phosphorylates Desnutrin/ATGL and hormone-sensitive lipase to regulate lipolysis and fatty acid oxidation within adipose tissue. Mol Cell Biol. 2016;36:1961–76.
7. Désert C, Duclos MJ, Blavy P, Lecerf F, Moreews F, Klopp C, et al. Transcriptome profiling of the feeding-to-fasting transition in chicken liver. BMC Genomics. 2008;9:611.
8. Bouvier-Muller J, Allain C, Tabouret G, Enjalbert F, Portes D, Noirot C, et al. Whole blood transcriptome analysis reveals potential competition in metabolic pathways between negative energy balance and response to inflammatory challenge. Sci Rep. 2017;7. https://doi.org/10.1038/s41598-017-02391-y.
9. Tall AR, Costet P, Wang N. Regulation and mechanisms of macrophage cholesterol efflux. J Clin Invest. 2002;110:7.
10. Rajavashisth TB, Dawson PA, William DL, Shackelford JE, Lebherz H, Lusis AJ. Structure, evolution, and regulation of chicken apolipoprotein A-I. J Biol Chem. 1987;262:7058–7065.
11. Freire-Regatillo A, Argente-Arizón P, Argente J, García-Segura LM, Chowen JA. Non-neuronal cells in the hypothalamic adaptation to metabolic signals. Front Endocrinol. 2017;8. https://doi.org/10.3389/fendo.2017.00051.
12. Kim WS, Weickert CS, Garner B. Role of ATP-binding cassette transporters in brain lipid transport and neurological disease. J Neurochem. 2008;104:1145–66.
13. Björkhem I, Meaney S. Brain cholesterol: long secret life behind a barrier. Arterioscler Thromb Vasc Biol. 2004;24:806–15.
14. Dietschy JM, Turley SD. Thematic review series: brain lipids. Cholesterol metabolism in the central nervous system during early development and in the mature animal. J Lipid Res. 2004;45:1375–97.
15. Pinto S, Roseberry AG, Hongyan L, Diano S, Shanabrough M, Cai X, et al. Rapid rewiring of Arcuate nucleus feeding circuits by Leptin. Science. 2004;304:110–5.
16. Nuzzaci D, Laderrière A, Lemoine A, Nédélec E, Pénicaud L, Rigault C, et al. Plasticity of the Melanocortin system: determinants and possible consequences on food intake. Front Endocrinol. 2015;6. https://doi.org/10.3389/fendo.2015.00143.
17. Buttgereit F, Brand MD. A hierarchy of ATP-consuming processes in mammalian cells. Biochem J. 1995;312:163–7.
18. Mitchell RW, Hatch GM. Fatty acid transport into the brain: of fatty acid fables and lipid tails. Prostaglandins. Prostaglandins Leukot Essent Fat Acids. 2011;85:293–302.
19. Mitchell RW, On NH, Del Bigio MR, Miller DW, Hatch GM. Fatty acid transport protein expression in human brain and potential role in fatty acid transport across human brain microvessel endothelial cells: fatty acid transport protein expression in human brain. J Neurochem. 2011;117:735–46.
20. Cedernaes J, Huang W, Ramsey KM, Waldeck N, Cheng L, Marcheva B, et al. Transcriptional Basis for Rhythmic Control of Hunger and Metabolism within the AgRP Neuron. Cell Metabolism. 2019;29:1078-1091.e5.
21. Dietrich MO, Horvath TL. Hypothalamic control of energy balance: insights into the role of synaptic plasticity. Trends Neurosci. 2013;36:65–73.
22. Di Marzo V, Matias I. Endocannabinoid control of food intake and energy balance. Nat Neurosci. 2005;8:585–9.
23. Bermudez-Silva FJ, Viveros MP, McPartland JM. Rodriguez de Fonseca F. the endocannabinoid system, eating behavior and energy homeostasis: the end or a new beginning? Pharmacol Biochem Behav. 2010;95:375–82.
24. Murataeva N, Straiker A, Mackie K. Parsing the players: 2-arachidonoylglycerol synthesis and degradation in the CNS: 2-AG synthesis and degradation in the CNS. Br J Pharmacol. 2014;171:1379–91.
25. Ezzili C, Otrubova K, Boger DL. Fatty acid amide signaling molecules. Bioorg Med Chem Lett. 2010;20:5959–68.
26. Bowen KJ, Kris-Etherton PM, Shearer GC, West SG, Reddivari L, Jones PJH. Oleic acid-derived oleoylethanolamide: a nutritional science perspective. Prog Lipid Res. 2017;67:1–15.
27. Guillou H, Zadravec D, Martin PGP, Jacobsson A. The key roles of elongases and desaturases in mammalian fatty acid metabolism: insights from transgenic mice. Prog Lipid Res. 2010;49:186–99.
28. de Antueno RJ, Knickle LC, Smith H, Elliot ML, Allen SJ, Nwaka S, et al. Activity of human Δ5 and Δ6 desaturases on multiple n-3 and n-6 polyunsaturated fatty acids. FEBS Lett. 2001;509:77–80.
29. Devane WA, Axelrod J. Enzymatic synthesis of anandamide, an endogenous ligand for the cannabinoid receptor, by brain membranes. Proc Natl Acad Sci. 1994;91:6698–701.
30. Jamshidi N, Taylor DA. Anandamide administration into the ventromedial hypothalamus stimulates appetite in rats. Br J Pharmacol. 2001;134:1151–4.
31. Ducheix S, Montagner A, Theodorou V, Ferrier L, Guillou H. The liver X receptor: a master regulator of the gut–liver axis and a target for non alcoholic fatty liver disease. Biochem Pharmacol. 2013;86:96–105.
32. Harizi H, Corcuff J-B, Gualde N. Arachidonic-acid-derived eicosanoids: roles in biology and immunopathology. Trends Mol Med. 2008;14:461–9.
33. Ohinata K, Yoshikawa M. Central prostaglandins in food intake regulation. Nutrition. 2008;24:798–801.
34. Kishore U, Reid KBM. C1q: structure, function, and receptors. Immunopharmacology. 2000;49:159–70.
35. Noris M, Remuzzi G. Overview of complement activation and regulation. Semin Nephrol. 2013;33:479–92.
36. Sjöberg AP, Trouw LA, Blom AM. Complement activation and inhibition: a delicate balance. Trends Immunol. 2009;30:83–90.
37. Ohinata K, Suetsugu K, Fujiwara Y, Yoshikawa M. Suppression of food intake by a complement C3a agonist [Trp5]-oryzatensin (5–9). Peptides. 2007;28:602–6.
38. Ohinata K, Takagi K, Biyajima K, Kaneko K, Miyamoto C, Asakawa A, et al. Complement C5a stimulates food intake via a prostaglandin D2- and

neuropeptide Y-dependent mechanism in mice. Prostaglandins Other Lipid Mediat. 2009;90:81–4.
39. Byerly MS, Petersen PS, Ramamurthy S, Seldin MM, Lei X, Provost E, et al. C1q/TNF-related protein 4 (CTRP4) is a unique secreted protein with two tandem C1q domains that functions in the hypothalamus to modulate food intake and body weight. J Biol Chem. 2014;289:4055–69.
40. Loftus TM. Reduced food intake and body weight in mice treated with fatty acid synthase inhibitors. Science. 2000;288:2379–81.
41. Obici S, Feng Z, Arduini A, Conti R, Rossetti L. Inhibition of hypothalamic carnitine palmitoyltransferase-1 decreases food intake and glucose production. Nat Med. 2003;9:756–61.
42. Minokoshi Y, Alquier T, Furukawa N, Kim Y-B, Lee A, Xue B, et al. AMP-kinase regulates food intake by responding to hormonal and nutrient signals in the hypothalamus. Nature. 2004;428:569–74.
43. López M, Varela L, Vázquez MJ, Rodríguez-Cuenca S, González CR, Velagapudi VR, et al. Hypothalamic AMPK and fatty acid metabolism mediate thyroid regulation of energy balance. Nat Med. 2010;16:1001–8.
44. Byerly TC, Kessler JW, Gous RM, Thomas OP. Feed requirements for egg production. Poult Sci. 1980;59:2500–7.
45. Bordas A, Merat P. Genetic variation and phenotypic correlations of food consumption of laying hens corrected for body weight and production. Br Poult Sci. 1981;22:25–33.
46. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2017. https://www.R-project.org/
47. Fox J, Weisberg S. An R companion to applied regression. Second. Thousand Oaks: Sage; 2011. http://socserv.socsci.mcmaster.ca/jfox/Books/Companion
48. Chiari Y, Galtier N. RNA extraction from sauropsids blood: evaluation and improvement of methods. Amphibia-Reptilia. 2011;32:136–9.
49. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
50. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
51. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC bioinformatics. 2011;12:323.
52. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11:R25.
53. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.
54. Zhou X, Lindsay H, Robinson MD. Robustly detecting differential expression in RNA sequencing data using observation weights. Nucleic Acids Res. 2014;42:e91.
55. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995;57:289–300.
56. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43:D447–52.
57. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.

Publisher's Note

III – Recherche de gènes *cis*-régulés impliqués dans l'EA par analyse d'expressions allèles-spécifiques

1. Recherche de gènes candidats causaux de la variation d'efficience alimentaire résiduelle, en combinant traces de sélection et recherche de gènes *cis*-régulés par analyse d'expression allèle-spécifique (article 5, en préparation)

a) contexte

Le contexte socio-économique de ce travail est le même que celui de l'article 3. Rappelons que l'amélioration de l'efficience alimentaire (EA) reste un défi pour accroitre la rentabilité des élevages et faire décroitre leur impact environnemental. L'aliment destiné aux animaux représente en effet 60% des coûts de production, et sa production est responsable d'une part importante des impacts environnementaux de l'élevage. L'enjeu scientifique est également similaire à celui de l'article 3, à savoir mieux comprendre la composante génétique de l'efficience alimentaire en utilisant le même modèle, les lignées R+ et R-, divergentes pour l'EA. Cependant, notre objectif est ici plus ambitieux puisque nous cherchons cette fois à identifier des gènes, voire des variants, causaux de la différence d'EA et de la différence des caractères associés à l'EA, qui ont été indirectement sélectionnés durant la sélection divergente sur 40 générations des lignées R+ et R-.

b) matériels et démarche

Nous avons utilisé les résultats d'un travail réalisé par ailleurs qui a permis de mettre en évidence des traces de sélection entre les lignées R+ et R-. Pour ce faire, des données de génotypage haute-densité ont été collectées sur 3 générations (générations 18, 25 et 40, 32 R+ et 32 R- par générations, pour un total de 192 animaux). Dans ces traces de sélection, qui sont des régions du génome dans lesquelles les différences de fréquences alléliques entre les lignées ne sont pas dues uniquement à la dérive génétique, devraient être localisés les variants causaux de la différence d'EA entre nos deux lignées.

Pour les gènes candidats causaux, nous avons élargi notre fenêtre d'observation autour des traces de 200 kb en 5' et en 3', car les variants *cis*-régulateurs (nécessairement dans la trace) peuvent être à distance du promoteur proximal du gène régulé. Le choix de 200 kb correspond

à un arrondi du troisième quartile de la distribution de la longueur des TAD détectés chez la poule par Foissac *et al.*[102] (2019), à savoir 160 kb. Les TAD[99] sont des régions transcriptionnellement actives dans lesquelles variants *cis*-régulateurs et gènes *cis*-régulés tendent à être localisés.

Nous avons d'abord cherché dans et à proximité de ces traces de sélection des gènes candidats causaux en utilisant leur annotation fonctionnelle, grâce aux fonctions connues dans la littérature et aux informations apportées par les observations phénotypiques chez des souris *knock-out*[364]. Ensuite, nous avons cherché des gènes candidats causaux à cause d'un variant affectant la structure du transcrit ou de la protéine associée, cette fois-ci uniquement dans les traces, et non plus à proximité (car le variant causal, nécessairement dans la trace, ne peut dans ce cas affecter qu'un gène dans la trace également). Nous avons ainsi cherché des variants des SNP ou des INDEL localisés dans les régions codantes des PCG et ayant un impact délétère sur la protéine codée. Concernant les LNC, bien qu'il soit difficile d'évaluer l'impact d'un variant dans ces gènes, nous avons relevé les INDEL de plus de 6 pb affectant les LNC dans ces traces. La valeur de 6 pb vient de l'étude de Kirk *et al.*[258] (2018) qui ont montré qu'il était possible de classer des LNC sur la base de leur composition en *k*-mers, avec $k = 6$ en particulier. Ces premiers travaux ont été entrepris à l'aide de données de DNA-seq collectées sur 10 animaux R+ et 9 animaux R-.

Enfin, nous avons cherché des gènes candidats causaux dont l'expression pourrait être régulée par un variant *cis*-régulateur localisé dans les traces de sélection. Pour cela, nous avons cherché les gènes, dans et à proximité des traces, présentant une expression allèle-spécifiques (ASE) dans des individus F_1 issus de croisements réciproques entre parents R+ et R- générés pour optimiser l'hétérozygotie des animaux. Pour les gènes ASE, la différence d'expression entre les allèles R+ et R- au sein des individus F_1 a été comparée à la différence d'expression au sein des individu F_0 des lignées R+ et R-. La recherche de ces gènes ASE *cis*-régulés s'est faite sur des échantillons de tissu adipeux, de sang, d'hypothalamus et de foie de 8 animaux F_1 âgés de 35 semaines et abattus à l'état nourri (6 individus pour l'hypothalamus).

c) résultats

La recherche des traces de sélection par l'équipe collaboratrice a mis en évidence 5 traces de tailles allant de 58 kb à 3 Mb sur 5 chromosomes différents.

Nous avons comparé ces traces de sélection avec des régions QTL pour des caractères en lien avec l'EA et observé que la trace de sélection sur le chromosome 6 correspond à une région

QTL pour la masse et le pourcentage de gras abdominal dans une population de poulets de chair[365].

Les 5 traces étendues contiennent 2 à 60 gènes, pour un total de 160 gènes, dont de 77 PCG et 78 LNC. D'abord, la recherche de gènes candidats positionnels et fonctionnels, c'est-à-dire dans ou à proximité d'une trace et avec une fonction connue en lien avec un phénotype divergent entre R+ et R-, nous a permis de mettre en évidence 17 PCG, dont deux également mis en évidence lors de la recherche de gènes *cis*-régulé : *TBC1D30* et *FABP4*. Le gène *TBC1D30* (à 157 kb de la trace du chromosome 6) a été identifié chez l'humain comme étant impliqué dans l'insulinémie et la sécrétion d'insuline par le pancréas[366,367] et *FABP4* (à 160 kb de la trace du chromosome 2) joue un rôle important dans le développement de maladies métaboliques (diabète de type 2, obésité)[368,369]. Ensuite, trois PCG présents dans les traces étaient affectés par un polymorphisme délétère dans les régions codantes : *LRRCC1* et *CNBG3* (tous deux dans la trace du chromosome 2) par un variant faux-sens (*missense*) prédit comme délétère, car l'acide aminé affecté est conservé à travers les espèces, et *SRRL* (dans la trace du chromosome 6) par un INDEL de 2 pb dans son dernier exon provoquant un décalage de lecture de 10% de la séquence codante. De plus, trois LNC étaient affectés par des INDEL, d'une taille allant jusqu'à 24 nt. Enfin, la stratégie qui vise à identifier des gènes candidats causaux *cis*-régulés par l'analyse conjointe des gènes ASE dans un des tissus « métaboliques » de la F_1 (tissu adipeux, foie et hypothalamus) et corroboré par le statut de DE entre R+ et R- a permis de mettre en évidence 9 PCG, et 1 LNC, soit 10 gènes en tout. En détail, nous avons identifié 6 gènes dans le tissu adipeux : *LLPH* (à 180 kb de la trace du chromosome 1), *CA13*, *FABP4* (dans et à proximité de la trace du chromosome 2, respectivement), *ANXA11*, *RET* (tous deux dans la trace du chromosome 6) et *HSBP1* (dans la trace du chromosome 11) ; 2 dans le foie, *MSRB3* et *TBC1D30* (dans et à proximité de la trace du chromosome 1, respectivement) et 2 dans l'hypothalamus, *COTL1* (à proximité de la trace du chromosome 11) et un LNC (INRAGALG00000008086 dans la trace du chromosome 2). Dans le sang, nous avons également identifié 8 PCG : *BMS1*, *MSRB3*, *OTUD6B*, *RMND1*, *SNX16*, *ZFAND1* et 2 LNC, gènes que nous n'avons pour l'instant pas étudiés en détails. Pour 6 des 10 gènes détectés dans le tissu adipeux, le foie et l'hypothalamus, l'expression était plus élevée chez le chromosome R- que chez le chromosome R+.

d) discussion

Nous avons fait ici le choix de ne discuter que le seul gène *FABP4* situé dans la trace de sélection du chromosome 2 car il est riche d'enseignements. Pour les autres gènes, on peut se référer à l'article en préparation ci-après.

FABP4, détecté dans le tissu adipeux et sur-exprimé par le chromosome R-, pourrait être un bon gène candidat positionnel et fonctionnel. D'abord, ce gène fait partie de la famille des *Fatty Acids Binding Proteins* (*FABP*), et est fortement exprimé dans les adipocytes, dans lesquels la protéine FABP4 facilite le trafic intracellulaire des lipides[370]. Mais FABP4 est également excrétée par ces cellules, et agit alors comme une adipokine[371]. Chez l'humain, *FABP4* est plus exprimé dans les adipocytes de patients obèses comparés à des patients sveltes, la protéine associée FABP4 est plus excrétée par ce tissu et sa concentration dans le sérum est positivement corrélée avec l'indice de masse corporelle[372] et associée avec l'insulino-résistance dans le diabète de type 2[373]. Enfin, et bien que les mécanismes semblent encore obscurs, *FABP4* semble impliqué dans le maintien de l'homéostasie du glucose, et dans des maladies métaboliques, dont le diabète[368]. Un examen plus précis de l'ASE pour chaque SNP composant le segment détecté par phASER nous montre cependant que ce ne sont que des SNP localisés dans le dernier intron du gène qui ont un $\log_2$ *allelic fold-change* réellement différent de 1, ce qui n'est pas le cas des autres SNP du gènes (dans les exons 2 et 4 sur 4 et dans les introns 1 et 2, voir aussi Figure 4B de l'article en préparation). L'existence d'un transcrit alternatif ne nous parait pas en être la cause. En effet, le nombre de *reads* associés à chaque SNP est le même pour tous les SNP dans les introns, qu'ils soient ASE ou non. Pour la même raison, nous ne pensons pas que cette observation soit due à l'existence d'un gène non-modélisé. L'absence d'ASE pour les SNP hétérozygotes des exons semble donc indiquer l'absence d'une *cis*-régulation de ce gène, pourtant excellent candidat positionnel et fonctionnel. Une analyse approfondie des SNP de ce gène, au-delà des SNP non-synonymes ayant des conséquences prédites délétères, montre l'existence d'un variant dans la lignée R+, qui entraine l'apparition d'un *missense* dans la protéine, et qui est associé à une moins forte affinité pour les acides-gras d'après Wang *et al*.[374] (2009). Ces auteurs montrent d'ailleurs que ce variant est présent uniquement chez des poulets de chair maigres[374], ce qui est cohérent avec le phénotype également maigre des R+. Ce gène reste donc un bon candidat pour sa fonction et la présence d'un variant affectant la structure et la fonction de la protéine associée et non d'un variant *cis*-régulateur. Se pose néanmoins la question des bornes des traces de sélection, puisque les variants présents dans *FABP4* (situé pour sa part hors de la trace) montrent bien des fréquences contrastées entre les deux lignées. D'une façon générale, se pose également le problème de la détection des variants non-

synonymes pouvant affecter sévèrement la protéine codée, autres que ceux prédits comme ayant un effet délétère. En effet, les variants *missense* annotés comme « tolérés » (comme le *missense* dont nous avons parlé dans *FABP4*) sont relativement nombreux, et affectent 19 gènes sur les 160 présents dans les traces. Le problème s'avère encore plus complexe avec les variants synonymes, qui peuvent également impacter sévèrement la protéine codée, notamment en affectant la vitesse de traduction[375,376].

e) conclusion

En conclusion, l'approche par détection de traces de sélection nous a permis de mettre en évidence 160 gènes candidats causaux de l'EA ou de caractères associés, sur les 52 075 que compte l'annotation étendue du génome de la poule. Même si cette première approche positionnelle est efficace dans la réduction du nombre de gènes candidats, il reste un important travail à fournir pour identifier les vrais gènes causaux parmi ces candidats positionnels. Dans ce contexte, nos trois stratégies de priorisation, basées sur (*i*) les fonctions connues des gènes, (*ii*) la présence de variants affectant la protéine codée ou l'ARN transcrit, (*iii*) la détection de *cis*-régulations potentielles à l'aide de l'étude de l'expression allèle-spécifique, et le croisement des résultats de ces trois stratégies, nous ont permis de grandement réduire cette liste de gènes potentiellement causaux en mettant en évidence 16 gènes candidats causaux (12 PCG et 4 LNC). La validation expérimentale du rôle de ces gènes par des approches de biologie moléculaire et cellulaire est maintenant l'étape suivante. Le choix du modèle cellulaire peut être guidé par les résultats d'expression allèle-spécifique et le *pattern* d'expression tissulaire des gènes candidats. En revanche le phénotype à suivre dans ces modèles cellulaires est difficile à définir. En effet, même si la détection de traces de sélection nous a permis de définir des régions d'intérêt avec assez peu de gènes (et en utilisant assez peu d'individus, à savoir 192 répartis en 3 générations), elle ne nous donne pas d'information sur les phénotypes associés à chacune d'entre elles. Il est donc difficile de savoir « quoi » chercher, c'est-à-dire quel phénotype chercher, *a fortiori* lorsque le phénotype est compliqué voire impossible à observer à l'échelle cellulaire (par exemple la prise alimentaire).

f) article en préparation

L'article associé à ces travaux est en cours de rédaction.

Il est reproduit ci-après dans son état d'avancement au moment de la fin de la rédaction du présent manuscrit.

Entre autres, la partie « *Material and Methods* » reste à être écrite, et les Figures 5 et 6 à être produites.

Titre prévisionnel : « Combined selective sweep mapping with coding SNP annotation and *cis*-eQTL analysis to identify candidate causative genes for phenotypic differences between RFI divergent laying hens »

Liste prévisionnelle des co-auteurs : Frédéric Jehl, Simon Boitard, Laetitia Lagoutte, Colette Désert, Christophe Klopp, Diane Esquerré, David Gourichon, Tatiana Zerjal, Sandrine Lagarrigue.

Combined Selective Sweep Mapping with Coding SNP Annotation and cis-eQTL Analysis to identify candidate causative genes for phenotypic differences between RFI divergent laying hens

Introduction

Feed efficiency (FE) represents the way animals use energy from their food for a productive purpose. It is a trait of great agronomical importance since feed costs represent 60 to 70% of the production costs in monogastrics, and the production of feed for livestock is responsible for a large share of livestock's deleterious impacts on the environment. Hence, pursuing FE comprehension and improvement remains a key challenge for the livestock sector. In this study, we used two experimental layer lines divergently selected since 1976 (i.e. during 40 generations) on the residual feed intake (RFI) [1], to investigate the causative genes of FE. RFI is a statistical built index used to estimate feed efficiency and is estimated as the difference between the observed feed intake (FI) and the predicted feed intake estimated considering body weight and egg production [2]. The animals of the efficient line have a negative RFI (R-) as their observed feed intake is smaller than the predicted FI. On the contrary, the animals from the inefficient line have a positive RFI (R+) as their observed feed intake is larger than the predicted FI. In addition to their difference in RFI, these lines are also diverging for FI (higher in R+ than in R-), abdominal fat weight (higher in R- than in R+), body temperature (higher in R+), or feather color (darker in R+), showing phenotypic correlation between RFI and multiple others traits. In terms of genetic correlations, Tixier-Boichard *et al*. have shown, in hens, low r_g between the RFI and the rectal temperature or shank length (r_g = 0.04 and 0.01), medium r_g between the RFI and wattle length (r_g = 0.19) and high r_g between the RFI and feed intake (r_g = 0.4) [3]. No correlation have been computed between body fat and RFI. Intense artificial selection of populations leads to the acceleration of evolutionary processes and often results in extreme phenotypes with associated changes across the genome. Using selective-sweep mapping, it is possible to identify the genomic regions where allele-frequencies have moved because of selection. In this paper, to detect candidate causative genes for these multiple traits that diverge between the lines, we focused on the five selective sweep regions detected (FDR of 15%) using the HapFLK method procedure [4, 5] well-adapted to the analysis of populations with small effective sizes, like the R+ and R- lines. This method focuses on the differences of haplotype frequencies between populations throughout several generations, and retains those for which the differentiation is too large to result from a neutral evolution model. Previous studies have shown that this strategy allows efficient control of the false-positive rate [4, 6]

even in the case of populations that underwent a bottleneck. To identify potential target genes in these selective sweep regions we used the enriched genome annotation we produced combined to original strategies that we will describe in the following paragraph. After analyzing these sweeps' content in terms of genes (protein coding genes – PCG and long non coding RNA genes – LNC) and variants (SNP and INDEL), we used two complementary strategies, as previously described in Roux *et al*. (2015) [5]. First we predicted the protein consequence of the SNP and INDEL located in the protein-coding genes (PCG) and studied in priority the cases predicted as deleterious and with "divergent" frequency between lines. These genes and the deleterious associated variants constitute a first set of candidate causative genes and variants which can be responsible for the trait variation through an alteration of the associated protein structure. Second, we focused on the candidate genes (called "eGenes") defined as having a *cis*-acting variant (called "*cis*-eQTL") in their regulatory regions responsible for a differentially expression between the R+ and R- genomes. Contrary to the first case, in this second case it is extremely difficult to identify these regulatory variants since they are considered to be numerous and poorly described. Hence, the strategy used consisted to study *cis*-regulations by allele-specific expression (ASE) using RNA-seq. Most studies interested in *cis*-regulations in farm species (cattle [7, 8], chicken [9, 10], pig [11]), provide an overview of ASE in some tissues and at some age but rarely used ASE strategy to identify candidate causal eGene by a *cis*-eQTL variant responsible for a particular trait [5, 12]. There have been however studies searching for *cis*-eQTL using genotype-phenotype association (with the phenotype being genes expression) by GWAS or linkage study, instead of ASE [13–15]. Here, we studied ASE in F_1 birds obtained from the crossing of one parent of each line to try to maximize the heterozygosity at the *cis*-regulatory region. Indeed, if a *cis*-regulatory variant of this region is the target site of selection it is expected to show allele fixation (or almost fixation) in one line, and the allelic frequency difference between the lines is expected to be large, although the variant targeted by the selection in one line is not necessarily expected to be absent from the other line. In F_1 animals the variant targeted by the selection should therefore be heterozygous in a certain number of birds. These two alleles of this potential *cis*-regulatory variant should induce an allele specific expression of the eGene in the F_1 birds and a differential expression between the two F_0 R- and R+ lines. An ASE in this work should not be in fact due to parental imprinting for two reasons: first, contrarily to mammalian species, no parental imprinting (*i.e*. monoallelic expression of autosomal genes from either the paternal or maternal allele) has been detected in chicken so far [16–18], and second, all F_1 birds come from reciprocal cross (i.e. half of them have a R+ mother and a R- father, the other half a R+ father and a R- mother), and we should

therefore be able to distinguish an ASE from an imprinting provided at least two individuals from the reciprocal cross are heterozygous. This strategy is dependent on a large number of variables, as the choice of the tissue(s), the physiological stage(s), the ages and the sex of the individual investigated, as it may impact the action of causal regulated gene(s) that is to be identified. We studied ASE in three tissues of F_1 birds which we consider as good "candidate tissues" considering the phenotypes that are divergent between lines: the adipose tissue, the key tissue for the energy storage, the liver, the key tissue of the lipid metabolism and energy homeostasis and the hypothalamus, an important regulatory center for feed intake and thermoregulation. RFI divergence selection is performed in adult animals at roughly 32 weeks of age, however divergence starts early in life, at around 8 weeks of age [19]. To investigate causative *cis*-regulations at an early age, the ASE analysis was also performed in F_1 birds of 10 weeks and 9 days of age. We also studied the blood, a circulating tissue that gather information from all over the organism in the adult F_1. These 4 tissues were collected on adult F_1 layers (35 weeks old) issued from parents that were divergent for RFI, FCR and body adiposity. As several metabolic traits diverged between lines at fasted condition [20] but not at fed conditions, we also studied the liver of adult F_1 layers which were feed deprived for 16 hours to detect potential changes in *cis*-regulated genes only expressed in this status. Because of the experimental cost the 2 ages and the fasting condition were analyzed only for the liver that we have prioritized because of its central role in energy homeostasis.

Results

Gene and SNP content of the five selective sweeps

Table 1: Overview of the 5 selective sweeps detected between the R+ and R- lines.

sweep					# genes				# SNP			# INDEL
chr.	start	end	length	line	total	PCG	LNC	other	F_1 RNA-seq	F_0 DNA-seq	%	F_0 DNA-seq
1	33 998 234	34 188 971	190 738	R-	6	1	5	0	864	1 945	95.7	214
2	122 145 397	125 147 671	3 002 275	R-	60	21	35	4	7 878	26 236	87.3	3881
6	4 778 250	6 200 576	1 422 327	R-	30	18	12	0	7 376	25 529	93.2	2567
7	31 616 846	31 674 988	58 143	R-	2	1	1	0	242	780	90.5	127
11	17 007 816	17 086 126	78 311	R+ & R-	3	2	1	0	581	898	85.2	98

"line" indicates the line in which one or more haplotypes seem to be fixed, or close to fixation.

"# genes": number of genes. PCG: protein coding genes. LNC: long non-coding RNA genes.

"# SNP": number of SNP. "%": percentage of F_1 RNA-seq SNPs present in F_0 DNA-seq.

"#INDEL": number of INDEL.

A previous work detected 5 selective sweeps (see Supplementary File 1), i.e., genomic regions in which the differences in allele frequencies between the lines were not compatible with a neutral evolution scenario, using high-density genotyping on 96 animals per line sampled at three different generations: the 18th, the 25th and the 40st generation of selection. Such an interval was shown by simulation to be the optimal solution to infer the presence of selection allowing to characterize the dynamic of selection (see also Material & Methods). Most of the sweeps had at least one haplotype fixed, or close to fixation, in the R- line except the sweep in chr. 11 for which one different haplotype was close to fixation in each line (see Supplementary File 1 and Table 1).

As shown in Table 1, these sweeps were located on chromosomes 1, 2, 6, 7 and 11, with length ranging from 58 kb (sweep chr. 7) to 3Mb (sweep chr. 2) and contained 2 to 60 genes, using an annotation of the chicken genome enriched in long non-coding RNA genes that we recently published, corresponding to a median of 42 genes per Mb (1st quartile – 3rd quartile: 34 – 52). More precisely, these sweeps contained 1 to 21 PCG and 1 to 35 LNC. In parallel, SNPs were detected in these sweeps, using either DNA-seq data (for the F_0 birds, that are the parents of the F_1 birds, see also Material & Methods) or RNA-seq data (for the F_1 birds). For the latter, we used RNA-seq data from 4 tissues (adipose tissue, hypothalamus, liver and blood) with the filters described in [Jehl *et al.*, article 3]. As expected according to [Jehl *et al.*, article 2], most ($\geq$ 85%) of the SNP detected by RNA-seq in the F_1 were also detected in the F_0 by DNA-seq. The sweeps contained between 780 and 26 236 SNP in the F_0 population, corresponding to a median of 11.5 SNP per kb (1st quartile – 3rd quartile: 10.2 – 13.4). Using QTL data from the Animal QTLdb database [21], we observed that the sweep chr. 6 included numerous QTL associated to abdominal fat weight and percentage [22] in two different populations of broiler chickens. These two broilers populations were a commercial dwarf population [23] or a cross between a commercial and a local Chinese breed [22], and are very different from the R+ and R- layers lines. The overlap between the sweep chr. 6 and QTL for fatness in two other populations suggests that this sweep is indeed a good candidate for harboring causative variants for fatness.

Strategy 1: Detection of candidate causative genes with a variant predicted to severely impact the structure of the associated RNA or protein

A

chr.	gene name	variant						Expr. [XX]			Expr. F_1		
		id	type	conseq.	nt or size	aa	freq R+ / R-	#tis.	top tis.	TPM	# tis.	top tis.	TPM
2	*LRRCC1*	rs312385904	SNP	missense	G>A	D>N	0 / 0.72	25	ovary	19.4	4	adip. tis.	2.7
2	*CNGB3*	rs13731105	SNP	missense	T>A	F>L	0 / 0.78	0	—	—	0	—	—
6	*SRRL*	—	INDEL	frameshift	at> aAAt	I > Kx	0.4 / 1	24	muscle	49	4	adip. tis.	2.8
2	INRAGALG00000008074	—	INDEL	×3	7, 20, 24	—	0, 0, 1 / 0.72, 0.72, 0.28	3	blood	1.32	2	blood	1.26
2	INRAGALG00000008062	—	INDEL	×3	6, 12, 17	—	1, 0, 0 / 0.28, 0.72, 0.72	6	blood	2	2	blood	2.3
6	INRAGALG00000014705	—	INDEL	×2	10, 12	—	0.65, 0.65 / 0, 0	7	blood	0.25	0	—	—

B

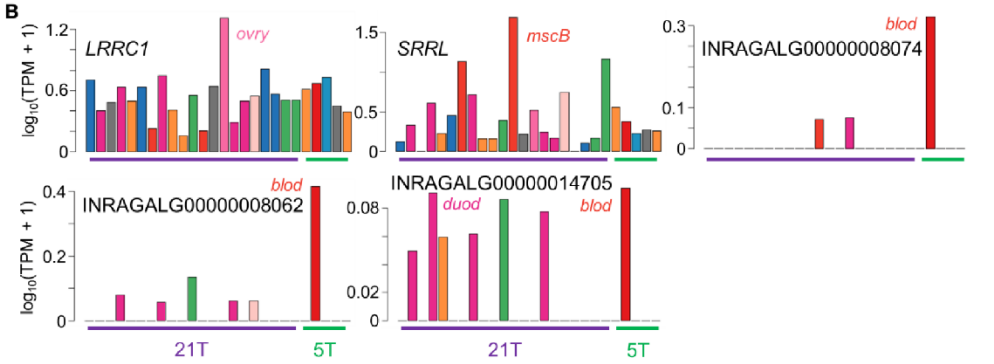


Figure 1: Overview of the genes affected by a variant with a potentially strong impact on the protein or the RNA transcript (for LNC). **A** Summary of key informations regarding the affected genes. "nt or size" indicates the coding nucleotides modified by the SNP or the INDEL, or the length of the INDEL for non-coding genes. When more than one INDEL affected a gene, their sizes are separated by a comma. "aa" indicates the modified amino-acids, "x" means that the second amino-acid was not determined by VEP. "freq R+/R-" provides the frequency of the SNP or the INDEL(s) in the R+ (left of the slash) and the R- (right of the slash). When more than one INDEL affected a gene, the frequencies are separated by a comma and given in the same order as their size. "#tis." number of tissues with expression, "top tis." tissue in which the gene's expression is the highest. "TPM" expression in TPM in the tissue in which the gene's expression is the highest. **B** Expression patterns of the 5 genes expressed in our data (*CNBG3* appears to be mostly expressed in eye parts which are not available, see main text). The order of the bars are, for the 21T: bursa of Fabricius, cecal tonsils, cerebellum, duodenum, adipose tissue around the gizzard, harderial gland, heart, ileum, kidney, liver, lung: lung, breast muscle, optical lobe, ovary, pancreas, proventriculus, skin, spleen, thymus, thyroid gland, trachea, and for the 5T: adipose tissue, blood, embryo, hypothalamus and liver. blod: blood, duod: duodenum, mscB: breast muscle, ovry: ovary.

Detection of 3 PCG with deleterious variants impacting the associated protein sequence – Since the selective sweeps should contain genes and variants involved in the difference in feed efficiency between the lines, we first screened for variants which were predicted to affect the protein in a deleterious way (see Material & Methods) and with contrasted frequencies between the two lines (see Table 2). Using DNA-seq data, we found 2 genes affected by a SNP inducing a missense and predicted as deleterious by SIFT (SIFT scores of 0.03 and 0.04, respectively),

both located in the sweep on chr.2, on a total of 41 missense SNP affecting 10 genes in this sweep. The first one was rs312385904, a G>A polymorphism inducing the translation of an asparagine instead of an aspartic acid in LRRCC1. The alternative allele had a frequency of 0.72 in R- versus 0 in R+. The second one was rs13731105, a T>A polymorphism inducing the translation of a leucine instead of a phenylalanine in CNGB3. The alternative allele had a frequency of 0.78 in R- versus 0 in R+. In addition, we found in the sweep in chr. 6 one interesting insertion out of 6887 INDEL across the 5 sweeps: the rs1058815473 insertion. It is a two-base pair insertion of AA in the last exon of ENSGALG00000002638 (*a.k.a SRRL*), affecting the amino-acids 473 to 505 (i.e. the last 10% of the protein) with a frequency of 1 in R- versus 0.4 in R+. *LRRCC1* appears to be highly expressed in the ovary (Figure 1B, using [Jehl *et al*., article 1] data, composed of two datasets: one with 21 tissues – "21T", one with 5 tissues – "5T", with only the liver in common, hence a total of 25 tissues) and in the adipose tissue (using this work's expression data). *CNGB3* for its part was not expressed in any of the 25 tissues of [Jehl *et al.*, article 1] (Figure 1B), nor in any of the four tissues used in this work. *SRRL* is a chicken orthologue of human's *SRR* gene, and had its highest expression in the breast muscle (Figure 1B).

Detection of 3 LNC affected by large INDEL – LNC are important regulators of gene expression, but they are currently poorly annotated, and their mechanisms of action are generally unknown. The prediction of the effects of the SNP is impossible since the primary structure is poorly conserved between species. However, it seems that *k*-mer of size $k = 6$ in LNC are conserved across species and allows clustering together LNC [24]. We hypothesized that polymorphisms targeting *k*-mer may alter the function of the affected LNC, and decided to focus on the INDEL of size ≥ 6 present in the LNC in the sweeps. In addition, we selected only INDEL with highly contrasted frequencies, i.e., entirely fixed in one line and entirely absent from the other (see Table 2).

Without consideration of frequency between lines, in total on the 5 sweep genomic regions we found 6887 INDEL (Table 1), of these 363 were affecting the exons of 95 genes (71 LNC and 24 PCG) . These INDEL had a median size of 3 bp (Q1 – Q3: 2 – 5), and there were a median of 3 INDEL (Q1 – Q3: 1 – 6) per LNC and 2 (Q1 – Q3: 2 – 3) per PCG.

Among them, we found 8 INDELs that had a length ≥ 6 and contrasted frequencies between the lines, affecting 3 LNC in the sweeps on chr. 2 and 3. The LNC were INRAGALG00000008074 and INRAGALG00000008062 in the sweep on chr. 2 and INRAGALG00000014705 in the sweep on chr. 3. The first LNC was affected by a 24-nt, a 20-nt and a 7-nt insertion and had a maximum expression of 1.32 TPM in the blood (Figure 1B). INRAGALG00000008062 was

affected by a 17-nt insertion and a 12-nt and a 6-nt deletions and had a maximum expression of 2 TPM also in the blood(Figure 1B). Finally, INRAGALG00000014705 was affected by a 12-nt and a 10-nt insertion, but shows low-to-no expression among our data (Figure 1B). Notably, INRAGALG00000008074 was classified as divergent of CA3A, at 5,304 bp of distance.

Strategy 2: Detection of candidate causal eGenes, i.e. differentially expressed because of a cis-regulatory variant

A

		sweep chr. 2 ext.			sweep chr. 6 ext.			sweep chr. 1 ext.			sweep chr. 11 ext.			sweep chr. 7 ext.			Total
Length		3.4 Mb			1.8 Mb			590 kb			478 kb			458 kb			
# genes	PCG	32			20			7			13			5			77
	LNC	43			17			9			4			5			78
	MIR	2			0			0			0			0			2
	Others	3			0			0			0			0			3
	Total	**80**			**37**			**16**			**17**			**10**			**160**
# funct. annot.	PCG	7			7			2			1			—			17
	LNC	1			—			—			—			—			1
# tis. specific – 5T	PCG	—			—			—			1 (liver : *HNF4 β*)			—			1
	LNC	1 (adipose tis.)			3			—			—			1 (adipose tis.)			2
		adip	hypt	livr	adip	hypt	livr	adip	hypt	livr	adip	hypt	livr	adip	hypt	livr	Total (t) or means (m)
# expressed	PCG	25	26	22	16	16	14	6	7	5	12	11	10	3	3	3	t 65
	LNC	4	7	4	3	6	1	0	0	3	0	0	0	2	1	1	t 21
	% genes	78	81	69	80	80	70	86	100	71	92	85	77	60	60	60	m 77
	% genes	9	16	9	18	35	6	0	0	33	0	0	0	40	20	20	m 14
# analyzable	PCG	19	19	19	15	15	14	6	6	5	10	10	9	2	2	2	t 58
	LNC	3	6	3	1	3	0	0	0	1	0	0	0	0	1	1	t 13
	% expr.	76	73	86	94	94	100	100	86	100	83	91	90	67	67	67	m 85
	% expr.	75	86	75	33	50	0	0	0	33	0	0	0	0	100	100	m 37
ASE ≥ 1 sample	PCG	7	7	9	8	2	6	4	4	4	4	3	4	2	0	0	t 49
	LNC	0	3	1	1	0	0	0	0	0	0	0	0	0	1	0	t 9
ASE ≥ 2 samples	PCG	4	3	4	2	0	3	4	2	4	3	0	2	1	0	0	t 36
	LNC	0	2	0	1	0	0	0	0	0	0	1	0	0	1	0	t 5
	% anal.	21	16	21	13	0	21	67	33	80	30	0	22	50	0	0	m 25
	% anal.	0	33	0	100	0	0	0	0	0	0	0	0	0	100	0	m 16
ASE ≥ 2 samples and DE between lines	PCG	2	1	—	2	—	—	1	—	3	—	—	—	—	—	—	t 9
	LNC	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	t 1

B

Sweep	line fixed	Gene name	Tissue	TPM F_1	log_2aFC	TPM F_0	log_2FC	# Het	# ASE	higher aFC	MGI
chr. 1 (ext.)	R-	LLPH	adipose tis.	110	0.51	67	0.35	8	2	R-	
chr. 1 (int)	R-	MSRB3	liver	18	-0.74	38	-0.71	8	3	R+	
chr. 1 (ext.)	R-	TBC1D30	liver	2.9	-1.31	3.7	-0.62	8	2	R+	✓
chr. 2 (int)	R-	CA13	adipose tis.	5.4	-1.45	4.2	-0.9	8	6	R+	
chr. 2 (ext.)	R-	FABP4	adipose tis.	63666	0.33	19899	0.33	8	7	R-	✓
chr. 2 (int)	R-	INRAGALG00000008086	hypoth.	2.7	2.41	2.8	1.69	6	5	R-	
chr. 6 (int)	R-	RET	adipose tis.	108	-0.67	188	-0.64	2	2	R+	
chr. 6 (int)	R-	ANXA11	adipose tis.	189	0.31	264	0.33	2	2	R-	
chr. 11 (int)	R+ & R-	HSBP1	adipose tis.	58	0.53	74	0.18	8	2	R-	
chr. 11 (ext.)	R+ & R-	COTL1	hypoth.	218	0.18	289	0.08	6	3	R-	

C

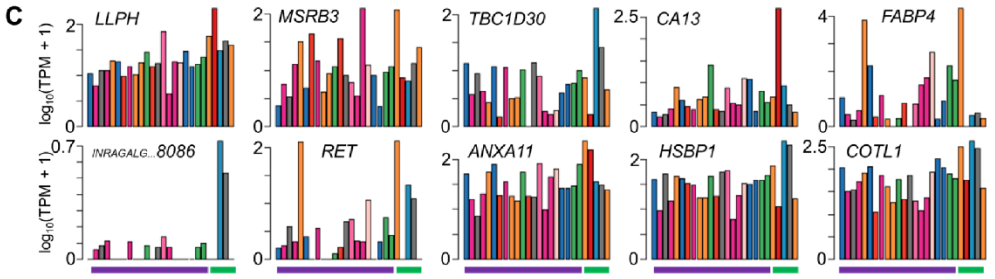


Figure 2 : Overview of the extended sweeps content and of the 10 candidate causative genes, with key features about the genes. A *Top*: summary of the extended sweeps and their content. The number of genes with a functional annotation as defined in Material & Methods are indicated, as well as the number of tissue-specific genes in the 5T dataset. *Bottom*: details by sweeps and tissues of the

expressed PCG and LNC (in number and percentage of total), of the analyzable PCG and LNC (in number and percentage of expressed), of the PCG and LNC ASE in one or more sample (in number), in 2 or more samples (in number and percentage of analyzable) and of PCG and LNC ASE in 2 or more samples and differentially expressed between lines, corresponding to the candidate causative genes. The right-most column provides totals (top part, and rows indicated by "t" in the bottom part) or means (bottom part, rows indicated by an "m") of the rows content. For the totals, the number of unique genes are given. **B** Summary of the 10 candidate causative genes. Sweep: localization relative to the sweep (int: within the sweep, ext: outside of the sweep). "Line fixed": line in which an haplotype of the sweep is fixed or close to fixation. "Tissue": tissue in which the gene is ASE. "TPM F_x": expression of gene in the F_1's or F_0's tissue (in TPM). "log_2aFC": log_2 allelic fold-change between the R- chromosome and the R+ chromosome. "log_2FC": log_2 fold-change between the R- line and the R+ line. "#Het": number of heterozygous samples. "#ASE": number of ASE samples. "higher aFC": line origin of the chromosome from which the gene has the highest expression. "MGI": indicates whether a term related to the lines phenotype was found in MGI. **C** Expression of the genes of panel B across the 25 tissues of [article 1]. The bars order is the same as in Figure 1. Purple lines indicate the tissues of the 21T dataset and green lines those of the 5T dataset.

Here our goal is to detect genes which expression differ between R+ and R- due to *cis*-regulation sites. Even though the *cis*-regulatory variants are all supposed to be within a sweep, it is possible that the regulated gene is out of the sweep. Indeed, *cis*-regulatory variants can be in an enhancer or a distal promoter region, which are by definition distant from the proximal promoter of the regulated genes. The GTEx consortium recently showed that pairs of cis-regulatory variants and regulated genes are significantly enriched for being in the same topologically associated domain (TAD) [25], consistent with the current model in which TADs act as scaffolds for enhancer-promoter interactions [26]. In mouse, these TADs are described as having a median size of 880kb, with finer domains (called "sub-TADs") with a median size of 185 kb [27]. In order to account for these TADs that gather enhancer-promoter regulatory elements, we analyzed all the genes at 200kb from either side of the sweeps. This value corresponds to the rounding up of the third quartile of the length of the TADs (160 kb) detected in chicken by Foissac *et al.* (2018) [28]. In addition, this size corresponds to the size of genomic block in linkage disequilibrium ($D > 0.3$) for the Rhode Island Red line [29].

First selection of positional and functional candidate

The objective in this section is to provide a first selection of positional and functional candidate causative genes, based on two criteria. The first one is the tissue-specificity (TS) using two independent data sets of 21 tissues and 5 tissues [Jehl *et al.*, article 1]. Genes with a TS in a

tissue related to the traits of interest could be a good candidate. The second criterion consist in using the MGI phenotype database to find a potential association between the gene name and some traits or diseases related to our R+ and R- models, characterized for example by an insulin-resistance or an hepatic steatosis (see Material & Methods for the terms used). For the LNC, we used the terms associated to the closest PCG as performed by several authors [28], considering that the members of PCG:LNC couples are susceptible to have a common function when they are in close proximity. Consistently, we recently showed in chicken [Jehl *et al.*, article 3, part regarding the LNC] using RNA-seq from the R+ and R- lines that when a LNC was DE between the lines, the closest PCG tended to be also DE more often in a LNC:PCG couple than in PCG:PCG couple, suggesting a common function for the LNC:PCG pair due to the proximity. For the LNC, we added information regarding the closest PCG, the configuration, and the distance that separates the LNC from the PCG pair and other features, etc. (see Supplementary Tables S1 for the PCG and S2 for LNC).

Regarding the terms, we found 17 PCG that had an associated phenotype relevant with the differences between our lines (see Material and Methods), and one LNC, INRAGALG00000008055, which was divergent at 11,105 bp of *FABP4*. Indeed, *FABP4* was associated to numerous relevant MGI terms such as "insulin resistance" or "impaired lipolysis".

Regarding the tissue expression, among the 160 genes in the 5 sweeps, 118 (74%) were expressed in at least one of the 25 tissues, of which 71 PCG and 45 LNC. Among the latter, 20 and 12, had an expression ≥ 0.5 and ≥ 1 TPM, respectively, which is to be taken into account for further experimental validation. Indeed, genes with an expression greater than 0.5 TPM are easier to detect with standard RT-qPCR. Among the 118 expressed genes, 21 genes (3 PCG and 18 LNC) were considered as tissue-specific (tau-score ≥ 0.95, [Jehl *et al.*, article 1]) using the "21T" dataset and 16 genes (3 PCG and 13 LNC) were specific in the "5T" dataset. 1 PCG was specific of one of the tissues of interest in relation to our phenotypes (see list in Material & Methods) and 5 LNC. In the liver, the only tissue-specific PCG was HNF4β, and INRAGALG00000014694 was the only tissue-specific LNC. In the adipose tissue were 3 tissue-specific LNC, INRAGALG00000008053 (TPM $\geq$ 0.7), INRAGALG00000014690 (TPM $\geq$ 3.4) and INRAGALG00000015786 (TPM $\geq$ 0.26) and in the hypothalamus was one LNC, ALDBGALG0000004897 (TPM $\geq$ 0.16).

ASE analysis in the 4 tissues of the 35 week-old F_1 birds slaughtered in a fed status

We then analyzed ASE in the 3 metabolic tissues and the blood of the 35 week-old F_1 birds slaughtered in a fed status and for which the expression of the F_0 parents was also available.

This allowed to compare for the genes presenting an allele specific expression, the fold change of ASE in the F1 with their DE fold chance between the F_0 lines, considering that it is expected to be DE between the F0 lines.

In the F_1, in median 78% (Q1 – Q3: 69 – 83) of the PCG and 9% of the LNC (Q1 – Q3: 0 – 20) were expressed in at least one tissue. A gene can only be analyzed if there was at least one heterozygous SNP with ≥ 10 reads for at least one allele. Using these criteria on the genes expressed in at least one tissue, a median of 86% for the PCG (Q1 – Q3: 75 – 94) and 63% of the LNC (Q1 – Q3: 0 – 75) were analyzable among the expressed genes. We considered a gene to show an ASE if an ASE was detected in at least 2 samples among the 8 samples available, since the sweep haplotypes and SNP are not totally contrasted between lines (for example, fixed in one and absent in the other line, see Supplementary Figure 1). With this threshold, 36 of the 58 analyzable PCG and 5 of the 13 analyzable LNC were considered as ASE in at least one tissue. Among the 41 analyzable genes with ASE in ≥ 2 samples within at least one tissue of the F_1 birds, we found a total of 10 unique genes that were both ASE and DE between lines in the same tissue, 6 genes in one of the 5 sweep and 4 genes within 200kb of the extremities of the sweeps. For the 10 genes, we observed a concordance between the line with the highest ASE in F_1 birds and the line with the highest expression in the DE analysis in F_0 lines. These 10 genes are composed of 9 PCG and 1 LNC (with TPM = 2.7). The R- chromosome derived portions had higher expression for 6 out of the 10 genes (*LLPH*, *FABP4*, INRAGALG00000008063, *ANXA11*, *HSBP1* and *COTL1*), and the R+ for the 4 remaining genes (*RET*, *CA13* and *MSRB3, TBC1D30*). The DE fold-change between the two lines and the ASE fold-change within the F_1 between the R- and R+ chromosome are given in Figure 2B as well as their mean expression. As we can see, in the 5 genes that were ASE in more than 2 samples, all the samples have a FC in the same orientation whatever the reciprocal parental crossing, excluding therefore an ASE due to an eventual imprinting event (which does not seem to occur in chicken [10]). Thereafter, we will use the term "candidate eGenes" to address these 10 genes (9 PCG and 1 LNC) that are in or at proximity of a sweep.

We added in Figure 2C, the tissue expression of these 10 samples. Most of them are expressed in all the tissues, although at variable levels. In particular, *FABP4* shows a high expression level in the adipose tissues (around the gizzard and abdominal), as well as *RET*, although at a lower level. Only INRAGALG00000008086 was not expressed in all tissues, but had an expression of 4 TPM (i.e., 0.7 in $\log_{10}$[TPM + 1]) in the embryo. In addition, using the MGI data base, we found 2 genes with phenotypes related to our lines: *FABP4* associated with "insulin resistance" and "increased fatty acid level" and *TBC1D30* associated with "abnormal lean body mass". It

is noteworthy that some of these genes are not well studied in the literature: *TBC1D30* has 2 associated publications and *LLPH* has 3.

ASE analysis in the liver of younger F_1 birds or in feed deprived F1 birds

In order to study ASE in other physiological stages and conditions, we collected liver samples from F_1 individuals at two additional ages: 10 weeks and 9 days. We found 2 ASE genes at 10 weeks for 8 and 2 F_1 birds: *MSRB3* and *HNF4β,* respectively, reinforcing the status of candidate gene for *MSRB3.* At 9 days, we found 3 ASE genes: *MSRB3*, *TBC1D30* and *ENSGALG00000009879*, a 1-to-many orthologues of human *TMBIM4*.

In addition, we collected liver samples from 35 weeks old F_1 birds slaughtered at the fasted status. We found 7 ASE genes, of which *MSRB3* in 4 F_1 birds, as well as *TMEM254* (3 birds), *LLPH*, *WWP1*, *TMEM55A*, *BMS1* and *CISD1* (2 birds for all of them). These results reinforce the position of *MSRB3* among the best candidate genes, a gene identified only on the basis of ASE analysis, which function could be linked to some of the traits of interest as ossification and adipose tissue development, as shown previously in cattle [30].

Focus on some candidate genes due to their cis-regulation

We detailed the genomic configuration of the candidate causative genes in the sweeps in chromosomes 1 and 2 (representing 6 genes out of 10) in the Figures 3 and 4. Below the configuration, the $\log_2$ allelic fold-changes of each SNP are provided, as well as the number of reads associated to them. The aggregated result of the $\log_2$ allelic fold-change at the gene scale is finally provided.

Sweep on chr. 1, with haplotype almost fixed in R- (Figure 3):

We detected 3 candidate causative genes, *TBC1D30*, *MSRB3* and *LLPH* (Figure 3A, B and C, respectively)

Sweep on chr. 2 with haplotype almost fixed in R- (Figure 4):

We detected 3 candidate causative genes as well, *FABP4*, *CA13* and INRAGALG00000008086 (Figure 4A, B and C, respectively)

Sweep on chr. 6 with haplotype almost fixed in R-:

We detected 2 candidate causative genes, *RET* and *ANXA11* (Figure to be done)

Sweep on chr. 11 with haplotype almost fixed in R-:

We detected 2 candidate causative genes, *HSBP1* and *COTL1* (Figure to be done)

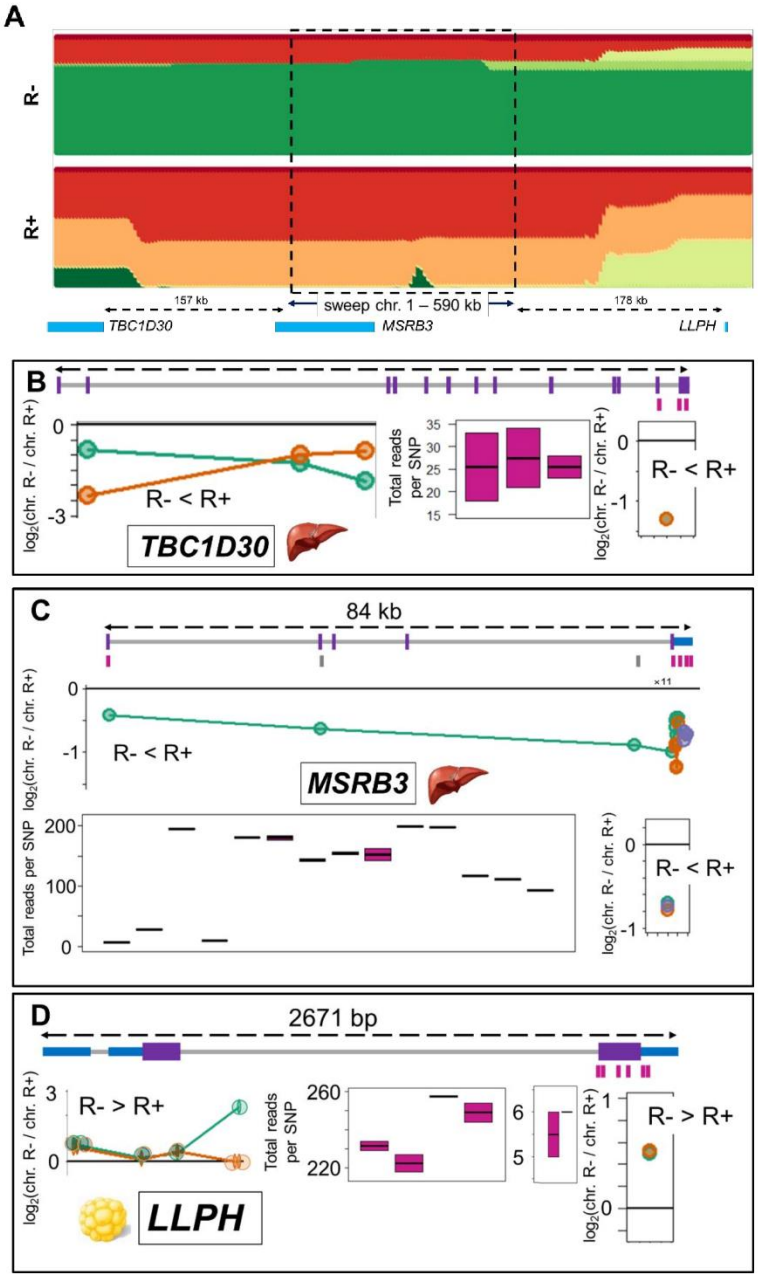


Figure 3: Candidate causative genes in the sweep in chr. 1. (**A**) haploFLK results for the sweep in chr. 1 and localization of the 3 candidate genes. (**B**) *TBC1D10* and (**C**) *MSRB3* in the liver. (**D**) LLPH in the adipose tissue. The gene model is given on top with coding exons (purple) the UTR (blue) and introns

(grey). The log_2(allelic Fold-change) is given for each SNP in every sample where it is available, and the resulting aFC for the whole haplotype is presented to the right. The boxplots represent the number of reads associated to each SNP in the exons (purple) and the introns (grey).

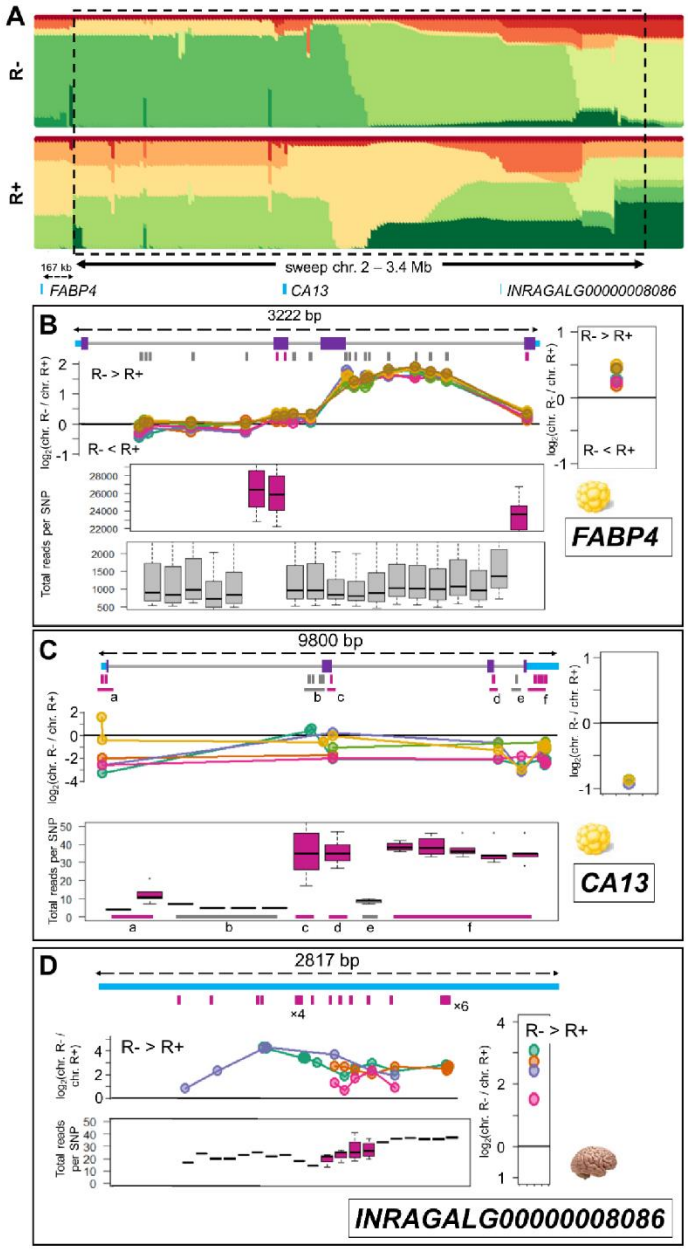


Figure 4: Candidate causative genes in the sweep in chr. 2. (**A**) haploFLK results for the sweep in chr.2 and localization of the 3 candidate genes. (**B**) *FABP4* and (**C**) *CA13* in the adipose tissue. (**D**) INRAGALG00000008086 in the hypothalamus. The gene model is given on top with coding exons (purple) the UTR (blue) and introns (grey). The log_2(allelic Fold-change) is given for each SNP in every

sample where it is available, and the resulting aFC for the whole haplotype is presented to the right. The boxplots represent the number of reads associated to each SNP in the exons (purple) and the introns (grey).

Discussion

A first set of 6 candidate causative genes due to the presence of a variant polymorphism potentially impacting the structure of the coded protein or transcribed RNA

These six genes were located in the two largest sweeps, the on chromosomes 2 and 6. The 3 PCG with potential structural variation did not present relevant MGI phenotype or disease term with respect to the differences between the lines. *LRRCC1* (a.k.a CLERC), is a centrosomal protein important for spindle pole integrity [31] and is affected by a missense predicted as deleterious. It was expressed in the ovary of the 21T dataset and the adipose tissue of the F_1 line (mean TPM = 2.7) making it an interesting candidate since the adipose tissue shows a strong divergence in mass between the two lines. *CNGB3* for its part seems to be involved in human in a rare autosomal disease called congenital achromatopsia, a condition also known as total color blindness [32]. It was not expressed in the 21T nor in the 5T tissue datasets. However, according to the GTEx data portal (in which there are no expression data related to eye but 53 available tissues) [33], this gene is mostly expressed in the testis (median TPM= 2.8), and to some extent, the fallopian tube (median TPM = 0.9). *SRRL* is a chicken ortholog of human's *SRR* gene that had its highest expression in the adipose tissue (mean TPM = 2.8). This gene was affected by a short insertion of two nucleotide, inducing a frameshift in the CDS, and most likely the translation of different amino-acids than the normal *SRRL*. However, since this insertion is located in the last exon and affects approximately the last 10% of the protein, it is likely that it only has a weak effect on the gene's function. The role of *SRRL* is not known, but the *SRR* seems to be involved in different brain functions, and although some behavioral differences were observed between lines (the R- being more fearful and less active than the R+), it is hard to make a link with any precise trait. The LNC genes are even more difficult to interpret, since less than 1% of the LNC have been studied and have a known function. We choose to signal the LNC genes that were affected by relatively large INDEL, since it appears that only small primary sequences were conserved and therefore are supposed to be important for LNC functions [24]. Among the three LNC, INRAGALG00000008074 and INRAGALG00000008062 had their highest expression in the blood, and both were also expressed in the pancreas. The latter was also expressed in the duodenum and the ileum.

INRAGALG00000014705 for its part had similar expression in the blood and the duodenum, and was also expressed in the pancreas and the ileum. Since the R- line show an insulin-resistant like phenotype, the pancreas might be a tissue to be studied in relation with this phenotype.

A second set of two candidate causative genes due to their function related to the traits or biological processes divergent between lines

We found 17 genes with relevant MGI terms. Among them, *FABP4* particularly retained our attention. *FABP4* is a member of the *Fatty Acids Binding Proteins* (FABP) and is highly expressed in adipocytes. The protein FABP4 plays an important role in the intracellular solubilization and trafficking of lipids [34], and is also secreted into the plasma, in which it plays a role as an adipokine [35]. *FABP4* plays important roles in the development of metabolic diseases such as obesity, type II diabetes. Indeed, *FABP4*-null mice are protected against the development of these metabolic diseases [36], and CRISPR interference of *FABP4* in obese mice induces body weight loss and inflammation, insulin resistance and hepatic steatosis diminution [37]. The high expression of *FAPB4* in R- can be related to this line's phenotype characterized by a hepatic steatosis and an insulin resistance, hence making this gene an interesting candidate causative gene to the R- phenotype because of its function and its position in the sweep chr. 2 fixed in R- at ~170 kb in 5' of the sweep, Figure 3A), and its function.

TBC1D30 was also part of the 17-genes list with relevant MGI terms ("abnormal lean body mass"). This gene was identified for bearing a low-frequency coding variants associated with fasting proinsulin or insulinogenic index in human [38]. In addition, in the pancreas, the diminution of *TBC1D30* expression is associated with the onset of diabetes because of its role as regulator of exocytosis [39]. In human, *TBC1D30* appears to be expressed in most of the tissues from the GTEx project [40]. While it is difficult to draw a clear link between *TBC1D30* expression in the liver and the insulin release in the pancreas, the lower expression of this gene in the R- genome can be related to the insulin-resistant phenotype observed in the R- line; and makes it a good candidate causative gene to the R- phenotype because of its function and its position in the sweep chr. 1 fixed in R- at ~157 kb of the sweep.

A third list of 10 cis-regulated eGene as candidate causal genes.

FABP4 (Figure 3B) is an interesting positional and functional candidate gene in the sweep chr. 2 and it was also detected in adipose tissue as ASE in 7 out of 7 analyzable birds and significantly DE between lines with high consistent fold change for ASE (average $\log_2$

aFC = 0.33) and DE (log_2 FC = 0.33) with higher expression in R- genome. However, a closer examination of the SNP-by-SNP ASE status in Figure 3B shows that only the SNPs located in the last intron do show a strong DE between the two lines, while the other SNPs, and especially those in the exons do not. It is noteworthy that all the SNPs in the last intron show an ASE with similar log_2 fold-change and that they are well-spread in the intron, suggesting that this observation is not due to a technical error. This is probably not due to an unknown isoform of *FABP4*, since the number of reads associated to each SNP is similar for all the SNP in the introns, including those in the last intron. For the same reason, the observed ASE is probably not due to a new unmodelized gene that overlap these SNP. The reason why only the last intron of *FABP4* show this ASE is not clear and casts some doubts on its status as a candidate causative gene. Its function is however very consistent with the phenotypic divergences observed between the lines.

A more thorough analysis of these gene's variants in our lines, showed that it is affected by a missense variant, predicted as "tolerated" by SIFT and fixed in the R+ line. This variant has be found to be present in a broiler chicken line with a low abdominal fat content and to potentially affect *FABP4* lipid binding capacity [41]. These consistent results also make *FABP4* a candidate causative gene due to a variant potentially affecting the function of the coded protein. This also raises the question of how exactly the hapFLK tool determine sweeps boundaries, since the visual observation of the haplotypes frequencies in the sweep and its surroundings (Figure 3) does not show variations between the inside and the outside of the boundaries.

TBC1D30 is a good positional and functional candidate in the sweep on chr. 1, and it was also detected in the liver as ASE in two samples out of two analyzable ones, with an average expression of 2.9 TPM in the F_1. This gene was under-expressed by the R- genome compared to the R+ one, with an average log2 allelic FC of -1.31 and a log2 fold-change between lines of -0.62. This ASE observed in the liver raises the question of a potential *cis*-regulation in other tissues, such as the pancreas.

In the liver, MSRB3 shows an unexplained expression and regulation pattern between the lines

MSRB3 was the only gene within the sweep on chr. 1 to be ASE (*TBC1D30* was located in the vicinity of this sweep), and it was also ASE at the three ages and in both nutritional conditions. In detail, it was detected as ASE in the liver of 18 F_1 birds (3 fed, 35 weeks old birds, 4 feed deprived 35 weeks old birds, 3 9-days old birds and 8 10-weeks old birds), with a lower

expression in R- line. *MSRB3* was highly expressed in the liver of the fed 35 week-old birds, with an average TPM = 18, and had a $\log_2$ allelic fold-change of -0.71 and a fold-change of -0.74. It is a ubiquitous gene, with expression in all of the 25 tissues from the 21T and 5T datasets, and with top expression in the proventriculus. However, the functional link between this gene and the divergent traits of interest in our model is not obvious. It was not among the 17 genes with a relevant MGI term, since the KO mice for this gene are characterized by hearing impairment [42]. However, this gene is known to be critical for the response to endoplasmic reticulum (ER) stress [43]. ER stress is a situation in which there is an accumulation of misfolded proteins in the ER, due to the inhibition of different biochemical reactions in this cellular compartment (*e.g.* glycosylation, disulfide bond formation) [44]. This stress can lead to the accumulation of intracellular reactive oxygen species (ROS) for example. ER stress can appear when the cell protein production exceeds the ER's processing capacity [45]. Depletion of this gene is mammalian cell induces an increase in reactive oxygen species [46], and can induce cell death [47]. This cellular phenotype is not easy to link to the animal-level phenotype. Indeed, we observed in the liver of the R+ line (in which *MSRB3* is over-expressed) an over-expression of genes involved in numerous amino-acids catabolism (5 of the 10 most enriched KEGG terms) and under-expressed genes involved in "protein processing in endoplasmic reticulum" and "ribosome activity" (the top 2 and 3 KEGG terms). We can note that *MSRB3* was detected ASE in the liver of 18 F_1 birds with a lower expression in R- line, in which the candidate variant was selected since the sweep in chr. 1 haplotype was fixed in the R- This is puzzling since it suggests that the variant fixed in R- negatively regulates the expression of this seemingly important gene.

The roles of most of the cis-regulated PCG remain to be elucidated

As noted in results, only 3 publications are available on Pubmed regarding *LLPH*, there are only 4 articles on Pubmed citing this gene, one in the frame of a GWAS related to blood pressure, one because it is located in a deleted region in relation with a familial condition of short stature and the last one because it was among the most connected gene in cancer data. *RET* on the other hand is a well-studied gene, but there are no clear link between its functions and our phenotypes. *ANXA11* for its part appears to be a RNA granule-associated phosphoinositide-binding protein, which acts as a molecular tether between RNA granules and lysosomes. Variants in this gene have been associated with amyotrophic lateral sclerosis, through an impairment of RNA granule transport [48]. *HSBP1* is a heat-shock protein and the literature review remains to be done for *COTL1*. To finish, *CA13* was found to be differentially

expressed in the liver between to group of broiler chicken with contrasted feed efficiency (*Personal Communication*).

INRAGALG00000008086 is the only candidate causative LNC

Finally, one candidate eGene was a LNC, INRAGALG00000008086. This gene was located in the sweep in chromosome 2 and was ASE in the hypothalamus of 5 of the 35 week-old birds, out of 6 analyzable, in which it had an expression of 2.7 TPM. Its role is of course difficult to infer, especially since the PCG it is classified with (MMP16, divergent at approx. 300 bp) is not correlated with this LNC in any of our tissues. Interestingly, INRAGALG00000008086 was expressed at 4 TPM in the embryo of our 5T dataset, suggesting that this gene could have a role to play in the development.

In summary, we found in the sweeps 160 genes that are positional candidate causative of our trait of interest, out of more than 52 000 genes in our extended annotation. This shows the interest of the selective sweeps mapping approach to clearly define a first general set of candidates. Using different priorizations, based on *i*) the known function of the genes, *ii*) the presence of a variant potentially impacting the structure of the protein or of the RNA and *iii*) the detection of a potential *cis*-regulation using allele-specific expression analysis, we detected 16 genes (4 LNC and 12 PCG) that could be good candidate causative genes, some of them being detected by more than one priorization. The validation of the involvement of these genes requires molecular and cellular biology manipulation, but the determination of phenotype to look for constitutes an obstacle to such a project. Indeed, while the sweep approach helped us define regions of interest with a relatively low number of samples (192 birds analysed), it does not give information on the associated phenotypes. Since the selection on a complex trait usually affect other associated traits, contrarily to a QTL regions analysis.

Material & Methods

Remains to be written.

Bibliography

1. Bordas A, Merat P. Genetic variation and phenotypic correlations of food consumption of laying hens corrected for body weight and production. British Poultry Science. 1981;22:25–33.

2. Bordas A, Tixier-Boichard M, Merat P. Direct and correlated responses to divergent selection for residual food intake in Rhode island red laying hens. British Poultry Science. 1992;33:741–54.

3. Tixier-Boichard M, Boichard D, Groeneveld E, Bordas A. Restricted Maximum Likelihood Estimates of Genetic Parameters of Adult Male and Female Rhode Island Red Chickens Divergently Selected for Residual Feed Consumption. Poultry Science. 1995;74:1245–52.

4. Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B. Detecting Signatures of Selection Through Haplotype Differentiation Among Hierarchically Structured Populations. Genetics. 2013;193:929–41.

5. Roux P-F, Boitard S, Blum Y, Parks B, Montagner A, Mouisel E, et al. Combined QTL and selective sweep mappings with coding SNP annotation and cis-eQTL analysis revealed PARK2 and JAG2 as new candidate genes for adiposity regulation. G3: Genes, Genomes, Genetics. 2015;5:517–529.

6. Fariello M-I, Servin B, Tosser-Klopp G, Rupp R, Moreno C, International Sheep Genomics Consortium, et al. Selection Signatures in Worldwide Sheep Populations. PLoS ONE. 2014;9:e103813.

7. Chamberlain AJ, Vander Jagt CJ, Hayes BJ, Khansefid M, Marett LC, Millen CA, et al. Extensive variation between tissues in allele specific expression in an outbred mammal. BMC Genomics. 2015;16:993.

8. Guillocheau GM, El Hou A, Meersseman C, Esquerré D, Rebours E, Letaief R, et al. Survey of allele specific expression in bovine muscle. Sci Rep. 2019;9:4297.

9. Ren P, Deng F, Wang Y, Ran J, Li J, Yin L, et al. Genome-wide analysis of spatiotemporal allele-specific expression in F1 hybrids of meat- and egg-type chickens. Gene. 2020;747:144671.

10. Zhuo Z, Lamont SJ, Abasht B. RNA-Seq Analyses Identify Frequent Allele Specific Expression and No Evidence of Genomic Imprinting in Specific Embryonic Tissues of Chicken. Sci Rep. 2017;7:11944.

11. Maroilley T, Lemonnier G, Lecardonnel J, Esquerré D, Ramayo-Caldas Y, Mercat MJ, et al. Deciphering the genetic regulation of peripheral blood transcriptome in pigs through expression genome-wide association study and allele-specific expression analysis. BMC Genomics. 2017;18:967.

12. Khansefid M, Pryce JE, Bolormaa S, Chen Y, Millen CA, Chamberlain AJ, et al. Comparing allele specific expression and local expression quantitative trait loci and the influence of gene expression on complex trait variation in cattle. BMC Genomics. 2018;19:793.

13. Le Bihan-Duval E, Nadaf J, Berri C, Pitel F, Graulet B, Godet E, et al. Detection of a Cis eQTL Controlling BMCO1 Gene Expression Leads to the Identification of a QTG for Chicken Breast Meat Color. PLoS ONE. 2011;6:e14825.

14. Ponsuksili S, Murani E, Schwerin M, Schellander K, Wimmers K. Identification of expression QTL (eQTL) of genes expressed in porcine M. longissimus dorsi and associated with meat quality traits. BMC Genomics. 2010;11:572.

15. Liaubet L, Lobjois V, Faraut T, Tircazes A, Benne F, Iannuccelli N, et al. Genetic variability of transcript abundance in pig peri-mortem skeletal muscle: eQTL localized genes involved in stress response, cell death, muscle disorders and metabolism. BMC Genomics. 2011;12:548.

16. Wang Q, Mank JE, Li J, Yang N, Qu L. Allele-Specific Expression Analysis Does Not Support Sex Chromosome Inactivation on the Chicken Z Chromosome. Genome Biology and Evolution. 2017;9:619–26.

17. Frésard L, Leroux S, Servin B, Gourichon D, Dehais P, Cristobal MS, et al. Transcriptome-wide investigation of genomic imprinting in chicken. Nucleic Acids Research. 2014;42:3768–82.

18. Wang Q, Li K, Zhang D, Li J, Xu G, Zheng J, et al. Next-Generation Sequencing Techniques Reveal that Genomic Imprinting Is Absent in Day-Old Gallus gallus domesticus Brains. PLoS ONE. 2015;10:e0132345.

19. Bordas A, Minvielle F. Patterns of growth and feed intake in divergent lines of laying domestic fowl selected for residual feed consumption. Poultry Science. 1999;78:317–23.

20. Swennen Q, Verhulst P-J, Collin A, Bordas A, Verbeke K, Vansant G, et al. Further Investigations on the Role of Diet-Induced Thermogenesis in the Regulation of Feed Intake in Chickens: Comparison of Adult Cockerels of Lines Selected for High or Low Residual Feed Intake. Poultry Science. 2007;86:1960–71.

21. Hu Z-L, Park CA, Reecy JM. Building a livestock genetic and genomic information knowledgebase through integrative developments of Animal QTLdb and CorrDB. Nucleic Acids Research. 2019;47:D701–10.

22. Sun Y, Zhao G, Liu R, Zheng M, Hu Y, Wu D, et al. The identification of 14 new genes for meat quality traits in chicken using a genome-wide association study. BMC Genomics. 2013;14:458.

23. Ye S, Chen Z-T, Zheng R, Diao S, Teng J, Yuan X, et al. New Insights From Imputed Whole-Genome Sequence-Based Genome-Wide Association Analysis and Transcriptome Analysis: The Genetic Mechanisms Underlying Residual Feed Intake in Chickens. Front Genet. 2020;11:243.

24. Kirk JM, Kim SO, Inoue K, Smola MJ, Lee DM, Schertzer MD, et al. Functional classification of long non-coding RNAs by k-mer content. Nat Genet. 2018;50:1474–82.

25. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369:1318–30.

26. Andrey G, Mundlos S. The three-dimensional genome: regulating gene expression during pluripotency and development. Development. 2017;144:3646–58.

27. Szabo Q, Bantignies F, Cavalli G. Principles of genome folding into topologically associating domains. Sci Adv. 2019;5:eaaw1668.

28. Foissac S, Djebali S, Munyard K, Vialaneix N, Rau A, Muret K, et al. Multi-species annotation of transcriptome and chromatin structure in domesticated animals. BMC Biol. 2019;17:108.

29. Hérault F, Herry F, Varenne A, Burlot T, Picard–Druet D, Recoquillay J, et al. A linkage disequilibrium study in layers and broiler commercial chicken populations. World Congress on Genetics Applied to Livestock Production (WCGALP). 2018;:6.

30. Saatchi M, Schnabel RD, Taylor JF, Garrick DJ. Large-effect pleiotropic or closely linked QTL segregate within and across ten US cattle breeds. BMC Genomics. 2014;15:442.

31. Muto Y, Yoshioka T, Kimura M, Matsunami M, Saya H, Okano Y. An evolutionarily conserved leucine-rich repeat protein CLERC is a centrosomal protein required for spindle pole integrity. Cell Cycle. 2008;7:2738–48.

32. Maguire J, McKibbin M, Khan K, Kohl S, Ali M, McKeefry D. CNGB3 mutations cause severe rod dysfunction. Ophthalmic Genetics. 2018;39:108–14.

33. GTEx Portal. https://www.gtexportal.org/home/gene/CNGB3. Accessed 2 Nov 2020.

34. Thompson BR, Mazurkiewicz-Muñoz AM, Suttles J, Carter-Su C, Bernlohr DA. Interaction of Adipocyte Fatty Acid-binding Protein (AFABP) and JAK2: AFABP/aP2 AS A REGULATOR OF JAK2 SIGNALING. J Biol Chem. 2009;284:13473–80.

35. Villeneuve J, Bassaganyas L, Lepreux S, Chiritoiu M, Costet P, Ripoche J, et al. Unconventional secretion of FABP4 by endosomes and secretory lysosomes. Journal of Cell Biology. 2018;217:649–65.

36. Prentice KJ, Saksi J, Hotamisligil GS. Adipokine FABP4 integrates energy stores and counterregulatory metabolic responses. J Lipid Res. 2019;60:734–40.

37. Chung JY, Ain QU, Song Y, Yong S-B, Kim Y-H. Targeted delivery of CRISPR interference system against Fabp4 to white adipocytes ameliorates obesity, inflammation, hepatic steatosis, and insulin resistance. Genome Res. 2019;29:1442–52.

38. Huyghe JR. Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. Nature Genetics. 2013;45:7.

39. Kolic J, Beet L, Overby P, Cen HH, Panzhinskiy E, Ure DR, et al. Differential Effects of Voclosporin and Tacrolimus on Insulin Secretion From Human Islets. Endocrinology. 2020;161:bqaa162.

40. GTEx Portal. https://gtexportal.org/home/gene/TBC1D30. Accessed 5 Nov 2020.

41. Wang Q, Guan T, Li H, Bernlohr DA. A novel polymorphism in the chicken adipocyte fatty acid-binding protein gene (FABP4) that alters ligand-binding and correlates with fatness.

Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology. 2009;154:298–302.

42. Kim M-A, Cho H-J, Bae S-H, Lee B, Oh S-K, Kwon T-J, et al. Methionine Sulfoxide Reductase B3-Targeted In Utero Gene Therapy Rescues Hearing Function in a Mouse Model of Congenital Sensorineural Hearing Loss. Antioxidants & Redox Signaling. 2016;24:590–602.

43. Lim D-H, Han JY, Kim J-R, Lee YS, Kim H-Y. Methionine sulfoxide reductase B in the endoplasmic reticulum is critical for stress resistance and aging in Drosophila. Biochemical and Biophysical Research Communications. 2012;419:20–6.

44. Kwak G-H, Lim D-H, Han JY, Lee YS, Kim H-Y. Methionine sulfoxide reductase B3 protects from endoplasmic reticulum stress in Drosophila and in mammalian cells. Biochemical and Biophysical Research Communications. 2012;420:130–5.

45. Navid F, Colbert RA. Causes and consequences of endoplasmic reticulum stress in rheumatic disease. Nat Rev Rheumatol. 2017;13:25–40.

46. Kwak G-H, Kim KY, Kim H-Y. Methionine sulfoxide reductase B3 deficiency stimulates heme oxygenase-1 expression via ROS-dependent and Nrf2 activation pathways. Biochemical and Biophysical Research Communications. 2016;473:1033–8.

47. Kwak G-H, Kim T-H, Kim H-Y. Down-regulation of MsrB3 induces cancer cell apoptosis through reactive oxygen species production and intrinsic mitochondrial pathway activation. Biochemical and Biophysical Research Communications. 2017;483:468–74.

48. Liao Y-C, Fernandopulle MS, Wang G, Choi H, Hao L, Drerup CM, et al. RNA Granules Hitchhike on Lysosomes for Long-Distance Transport, Using Annexin A11 as a Molecular Tether. Cell. 2019;179:147-164.e20.

Discussion, perspectives

I – Gènes et efficience alimentaire, des liens fort complexes

Nos travaux avaient pour premier objectif de mieux comprendre les mécanismes impliqués dans la différence d'efficience alimentaire (EA) en étudiant des lignées modèles divergentes pour ce caractère par des approches associant différentes données *–omics.* Ces données constituées de transcriptomes multi-tissus et de lipidome et métabolome hépatique, ont permis de mettre en évidence les gènes et voies métaboliques associés à la sélection pour cette différence d'EA. On peut se demander si les différences mises en évidence sont des conséquences de la sélection pour l'EA, ou bien s'il s'agit de caractères indirectement sélectionnés. Cette question se pose particulièrement pour la masse de gras corporel et l'insulino-résistance qui sont différentiels entre les lignées et pour lesquels de nombreux gènes associés sont différentiellement exprimés entre lignées ou localisés dans les traces de sélection. Pour répondre en partie à cette question « cause–conséquence », il faudrait calculer les corrélations génétiques entre ces caractères, notamment avec la masse de gras abdominal, ou avec des paramètres mesurant la résistance à l'insuline. Toujours est-il que ces premiers résultats nous ont aidés dans la recherche de gènes candidats causaux présents dans les 5 régions identifiées comme étant sous sélection et caractérisant notre modèle : d'une part en précisant les fonctions biologiques différant entre R+ et R-, et donc potentiellement impactées par la sélection, par exemple la stéatose hépatique ou l'inflammation du tissu adipeux, et d'autre part en identifiant les gènes différentiellement exprimés entre ces lignées, qui ont pu être mis en regard avec ceux ayant une expression allèle spécifique (ASE) dans les individus F_1. En effet, le second objectif était d'analyser les 160 gènes candidats positionnels localisés dans les 5 traces de sélection identifiées dans nos lignées, en dressant la liste des meilleurs gènes candidats causaux de l'EA, en raison (*i*) d'une fonction clairement liée aux phénotypes divergeant entre lignées, (*ii*) d'un variant affectant la structure de la protéine ou pour les LNC de l'ARN produit ou (*iii*) d'un variant *cis*-régulateur entrainant une variation d'expression entre les deux génomes R+ et R-, dans les F_1 (gènes ASE) retrouvée dans les F_0 (gènes DE).

Comme nous l'avons vu à différentes reprises, les lignées R+ et R- ont été sélectionnées pendant 40 ans de façon divergente pour la prise alimentaire résiduelle (RFI). Rappelons que cette prise alimentaire résiduelle est calculée comme la différence entre la prise alimentaire observée et la prise alimentaire prédite par une régression multiple prenant en compte le poids du corps, sa variation et la production d'œufs. Cette sélection a donc agi sur un caractère complexe qui est lui-même la résultante de plusieurs caractères complexes. Ainsi, au sein même

d'une des deux lignées, différents individus peuvent présenter un même phénotype (efficient ou inefficient), mais présenter des différences au niveau de composantes du RFI prises en compte dans le modèle. Par ailleurs, la sélection sur un critère, peu importe sa complexité, impacte très souvent d'autres caractères qui sont génétiquement corrélés avec ce critère, et pour lesquels la relation physiologique avec le critère sélectionné est plus ou moins évidente. Une étude réalisée en 1995[334] montre dans ces lignées que la RFI est fortement corrélée génétiquement avec la prise alimentaire (corrélation génétique $r_g = 0.4$, donc prise alimentaire plus élevée chez les R+), moyennement corrélée avec la longueur des barbillons ($r_g = 0.19$, donc plus longs chez les R+ également), mais très peu avec la température rectale ($r_g = 0.04$) et la longueur des tarses ($r_g = 0.01$) toutes deux plus élevées chez les R+. En revanche la corrélation génétique avec la masse de gras abdominal, qui est également très contrastée entre les lignées (plus élevée chez les R-) n'est pas connue. Notons que cette association « efficience alimentaire – adiposité corporelle » observée après la sélection sur la RFI semble généralisable, mais pas le « signe » de l'association, qui est positif dans notre modèle : la lignée efficiente mange peu et est plus grasse. En effet chez le porc, entre des lignées divergentes pour la RFI, la lignée efficiente mange également moins mais présente un taux de gras corporel plus faible que la lignée inefficiente[377], de même que chez des bovins divergents pour la RFI[378,379], et chez des poulets de chair[380].

Cette sélection indirecte pour différents caractères complexes, conséquence de la sélection divergente imposée depuis plus de 40 ans pour l'EA rend mécaniquement difficile l'analyse du contenu des traces de sélection. En effet, des gènes avec des fonctions très diverses et exprimés dans des tissus variés peuvent être responsables d'une des divergences observées, et, peu, voire pas, de gènes ne sont actuellement connus comme étant impliqués dans un caractère aussi complexe que la RFI ou l'EA. Alors que l'approche « gène candidat positionnel » (car présent dans une trace de sélection) permet déjà de grandement réduire la complexité de l'analyse (ici 160 gènes positionnels parmi les ≃50 000 gènes présents dans le génome entier), la priorisation par la fonction, à savoir l'implication dans le caractère est plus compliquée.

Par ailleurs, eu égard aux différentes combinaisons de caractères permettant d'obtenir une RFI variable, on peut s'étonner que seules cinq traces de sélection aient été détectées (avec pour rappel un seuil de significativité $p_{FDR} \leq 0.15$). Comparons par exemple avec une étude de 2015 menée par notre équipe[381] sur des lignées de poulets de chair divergentes pour l'adiposité corporelle qui a donné lieu à la mise en évidence de 10 régions d'intérêt chevauchant à la fois

des traces de sélection et des régions QTL de ce caractère. Une autre étude de 2015[382] portant sur des poulets de deux lignées divergentes pour le poids du corps a pour sa part révélé 16 traces, avec un seuil de significativité $p_{FDR} \leq 0.20$. Dans notre cas, utiliser ce dernier seuil aurait révélé 7 traces au lieu de 5. Citons finalement une étude sur des lignées de souris sélectionnées de façon divergente sur le poids du corps depuis une cinquantaine d'années qui a révélé 18 régions QTL pour ce caractère[383].

Enfin, une dernière difficulté afférente à l'étude de ces deux lignées vient de l'absence d'une lignée « contrôle », c'est-à-dire dans laquelle la sélection pour la RFI n'aurait pas eu lieu, et à laquelle on pourrait comparer les lignées R+ et R- indépendamment. En l'absence de ce contrôle, il est difficile de déterminer si une différence observée lors des analyses différentielles entre les R+ et les R- pour les variables -*omics* vient d'événements se déroulant dans les deux lignées dans des directions opposées, ou au contraire se déroulant dans une lignée mais pas dans l'autre. Il apparait que c'est surtout ce dernier cas qui s'est produit, comme le montrent les fréquences haplotypiques dans les traces de sélection que nous avons présentées dans l'article 5 (en préparation), avec 4 traces (sur les chromosomes 1, 2, 6 et 7) ayant un haplotype presque fixé dans une lignée, la R-, une trace (sur le chromosome 11) pour laquelle il semblerait qu'il y ait eu une fixation « opposée » de deux haplotypes entre les lignées.

Les R+ et R- n'en restent pas moins de précieux modèles pour étudier les gènes candidats causaux de la variation de RFI mais aussi de la variation de caractères associés à la RFI, en particulier de l'adiposité corporelle (sous réserve d'une corrélation génétique significative avec la RFI). En effet, nous avons clairement démontré que le foie des R- présente tous les signes de la stéatose hépatique (*cf.* article 3). D'abord, la comparaison des transcriptomes hépatiques entre R- et R+ suggère fortement une synthèse *de novo* dans le foie des R-, corroborée par (*i*) l'analyse lipidomique montrant un niveau plus élevé de la quantité d'acides gras et de triglycérides et de la proportion d'acide gras mono-insaturés par rapport aux poly-insaturés, par (*ii*) l'analyse de coupe de foies montrant un nombre plus important de gouttelettes lipidiques dans les R- et enfin par (*iii*) l'observation visuelle lors de la dissection de foies en majorité blanchâtres dans la lignée R-. De même, nous avons montré que le tissu adipeux des R- présente des signes suggérant une inflammation, probablement liée à l'excès d'acides gras stockés en grande quantité dans ce tissu chez cette lignée. Une telle observation est classique chez les sujets obèses[384–386] et pourrait notamment être due à l'activation par les acides-gras en excès de la chaîne respiratoire mitochondriale, qui est productrice d'espèces réactives de l'oxygène[387]. Nous avons également observé des signes de remodelage de ce tissu, un processus observé en

lien avec l'obésité, lors duquel le tissu adipeux subi des modifications fonctionnelles et morphologiques qui lui permettent d'accommoder la quantité de lipides qui lui arrive[388,389].

La recherche de gènes responsables des différences entre lignées R+ et R- (article 5 en préparation) nous a permis d'identifier 16 bons candidats parmi les 160 gènes présents dans les traces. D'abord, 6 le sont en raison d'un variant prédit comme affectant sévèrement la structure de la protéine codée (pour 3 d'entre eux, tous trois des PCG) ou en raison de la présence d'un ou plusieurs larges INDEL dans le gène (pour les 3 autres, tous trois des LNC). Parmi ces 6 gènes 2 PCG sont exprimés dans tous les tissus pour lesquels nous disposons de données d'expression ou presque (respectivement 24 et 25 tissus sur les 25 utilisés dans l'article 1), et le troisième dans aucun de ces tissus. Ce troisième PCG semble en fait exprimé dans l'œil (pour lequel nous n'avions pas de données d'expression), et d'après les données d'expression humaines du projet GTEx (qui n'ont pas non plus de données sur l'œil), dans le testicule. Les 3 LNC ont pour leur part été détectés comme exprimés dans 3 à 7 tissus. En tous cas, aucun de ces gènes (ni les PCG, ni a fortiori les LNC) ne semble avoir de fonction connue en lien avec les phénotypes.

Ensuite, 10 gènes sont candidats en raison d'une *cis*-régulation observée grâce à l'ASE (9 PCG et 1 LNC) : 6 dans le tissu adipeux, 2 dans le foie et 2 dans l'hypothalamus. C'est dans ce dernier tissu qu'est d'ailleurs exprimé le seul LNC détecté comme ASE, avec TPM > 2.6, ce qui le place dans le top 25% des LNC les plus exprimés. Parmi les 9 PCG, 2 nous paraissent avoir des fonctions clairement en lien avec les différences entre R+ et R- : *TBC1D30* et *FABP4*. *TBC1D30* (ASE dans le foie) a été identifié chez l'humain comme étant impliqué dans l'insulinémie et la sécrétion d'insuline[366,367], et *FABP4* (ASE dans le tissu adipeux) joue un rôle important dans le développement de maladies métaboliques (diabète de type 2, obésité)[368,369]. En revanche, il est beaucoup plus compliqué de comprendre le rôle des 8 autres gènes dans les différences entre R+ et R-. À ce stade, il faudrait procéder à des analyses fonctionnelles, par exemple invalider ou sur-exprimer ces gènes dans des modèles cellulaires adaptés pour observer éventuellement un phénotype en lien avec les phénotypes d'intérêts. Ce dernier point nous montre justement les limites de cette approche.

En effet, ce travail nous a permis de constater à quel point la mise en évidence de gènes candidats causaux *cis*-régulés pouvait être complexe. De fait, la majorité des gènes causaux d'un caractère plus ou moins complexe reportés dans la littérature le sont en raison d'un variant affectant la protéine qu'ils codent, ce qui vient du fait que l'effet de ces variants est facile à prédire. La détection de *cis*-eQTL par ASE à l'aide de données RNA-seq ouvre à présent la

voie à l'étude des *cis*-régulations grâce au coût abordable de cette technologie et aux effectifs relativement faibles nécessaires à une telle analyse, quoiqu'ils doivent tout de même être issus de dispositifs expérimentaux appropriés (dans notre cas, des croisement réciproques d'animaux sélectionnés et très divergents). Dans le cadre de la recherche de gènes responsables de caractères d'intérêt, ces études ASE peuvent donc être combinées à des traces de sélection de populations divergentes, mais également à des régions QTL responsables de la variation d'un caractère d'intérêt. La limite de l'approche par traces de sélection par rapport aux régions QTL est qu'il n'y pas de lien entre la région analysée et les caractères d'intérêt. Comme on a pu le voir dans la partie III, cela rend plus difficile la recherche du ou des gènes candidats causaux, la compréhension de leur rôle dans le caractère complexe d'intérêt et plus encore leur validation dans des modèles cellulaires, qui doivent être choisis en fonction du phénotype à étudier. En effet, si l'ASE permet de « cibler » le tissu, ce qui constitue déjà une information utile pour choisir le modèle cellulaire le plus pertinent, les traces de sélection ne permettent pas de savoir quel phénotype étudier dans le modèle choisi. Ces limites tombent avec l'analyse de régions QTL qui sont justement détectés pour leur lien avec un phénotype.

Dans tous les cas, il est fondamental d'étudier l'ASE à l'aide de données venant de tissus liés au caractère d'intérêt, toute l'approche reposant sur l'étude de l'expression génique. De plus, par définition, il faut, pour pouvoir étudier un gène dans un tissu que celui-ci présente au moins un SNP hétérozygote dans ses régions exprimées. Ce dernier point ne semble cependant pas être un obstacle majeur. Nous avons en effet vu dans l'article 2 qu'en moyenne 72% des PCG exprimés à TPM ≥ 1 dans un tissu présentent au moins un SNP hétérozygote dans les régions exoniques, quoiqu'avec des variations entre populations (48% à 89%). En ce qui concerne l'ASE, la dernière limite vient de l'insensibilité du test binomial aux faibles valeurs d'expression (illustrée par la Figure 7, partie I). En effet, nous avons observé dans le cadre de l'analyse ASE que pour de faibles valeurs de comptage, le différentiel d'expression entre les deux allèles devait être d'autant plus élevé que les comptages associés aux deux allèles sont faibles pour être détecté comme significativement ASE. L'outil phASER justement, et contrairement à ASEReadCounter, cumule les *reads* à l'échelle d'haplotypes, ce qui permet de faire une « synthèse » du ratio d'expression à l'échelle du gène, puisque cet outil priorise le segment supporté par le plus de *reads* (sur lequel les haplotypes sont définis). En revanche on peut se demander s'il ne serait pas plus pertinent de sélectionner les SNP à prendre en compte au niveau des exons uniquement, puisque l'on sait grâce aux travaux des parties I et III (articles 2 et 5) que de nombreux SNP détectés par RNA-seq sont présents dans les introns mais

représentent des transcrits encore immatures, ce qui peut conduire à des résultats difficiles à interpréter, comme dans le cas de *FABP4*.

En ce qui concerne les variants affectant la structure de la protéine codée, nous avons procédé à une priorisation sur le critère de la gravité prédite de l'effet du variant (en particulier, son score SIFT) comme cela se fait couramment, le nombre de SNP non synonymes étant beaucoup plus important. En effet, nous avions identifié au total 77 variants « *missense* » (prédits ou non comme délétères) dans les traces, et avons donc dû faire un choix quant à ceux que nous avons analysés. Cependant, l'exemple du variant « missense » situé dans *FABP4*, prédit comme n'étant pas délétère mais dont il a été démontré qu'il altérait la capacité de la protéine FABP4 à lier les lipides[374] nous montre bien que cette priorisation, si elle est nécessaire pour étudier un nombre raisonnable des gènes et de variants, conduit à laisser de côté des gènes et de variants potentiellement pertinents. L'exemple ici est à nuancer, car *FABP4* se trouve hors d'une trace (167 kb en 5' de la trace du chromosome 2), et les variants qui s'y trouvent ne devraient donc pas être considérés comme candidats causaux. Cela dit, l'observation, peut-être naïve de notre part, n'étant pas spécialistes de la méthode, des fréquences haplotypiques dans la région nous conduit à nous interroger sur la manière dont les bornes des traces de sélection ont été choisies par hapFLK. En effet, ces fréquences ne nous semblent pas particulièrement différentes de celles « dans » la trace.

II – L'annotation des ARN longs non-codant demande encore de nombreux efforts

Notre travail sur l'extension de l'annotation de référence Ensembl avec des modèles de LNC a permis de combler quelques lacunes dans l'annotation du génome de cette espèce d'élevage, au moins en termes de nombre de modèles. La difficulté majeure d'un travail sur des LNC reste la compréhension de leurs fonctions, puisqu'aucun outil ne permet aujourd'hui de les inférer en utilisant uniquement des motifs présents dans la séquence, comme c'est le cas avec les PCG. Il est en effet possible d'associer une fonction aux PCG en se basant sur la similarité de séquence entre espèce ou au sein d'une espèce[340] ou bien sur les motifs présents dans la protéine codée[339], approches impossibles avec les LNC. Cela étant, le travail de Kirk *et al*.[258] (2018) pourrait ouvrir la voie à des annotations fonctionnelles basées sur les profils de *k*-mers (une forme de motifs, donc) présents dans les LNC. On peut penser qu'il pourrait être possible soit d'associer une fonction à un ou des *k*-mers, soit de regrouper les LNC sur la base de leur composition en *k*-mers et d'associer à chaque groupe la ou les fonctions des LNC connus qui en font partie. Ces approches basées sur les séquences et les motifs qu'elles présentent nécessitent néanmoins que les fonctions de suffisamment de LNC soient élucidées.

L'approche par la classification des LNC à l'aide du gène le plus proche, et en particulier le PCG, n'en semble pas moins intéressante. En effet, elle est d'abord soutenue par différents exemples trouvés dans la bibliographie montrant des régulations de PCG par des LNC en positions divergentes et *antisens*[281,282,285], ou bien par l'observation qu'un LNC accueillant un micro-ARN (MIR) est non seulement précurseur de ce MIR, mais a en plus des fonctions liées à celles du MIR[390].

Enfin, rappelons que le consortium GTEx a observé que 7.5% des variants *cis*-régulateurs pouvaient agir sur plusieurs gènes à la fois, et qu'un quart de ces variants régulant plusieurs gènes régulaient des PCG et des LNC (et un tiers régulait plusieurs PCG)[273]. Dans la mesure où ces régulations ont tendance à agir au sein d'un même TAD[99], on comprend que le LNC a de bonnes chances de subir la même *cis*-régulation qu'un PCG proche, voire que le PCG le plus proche de lui. C'est d'ailleurs ce qui a été observé : lorsque les auteurs ont observé grâce à l'ASE la *cis*-régulation d'un LNC, les PCG autour de lui tendaient à être également ASE (donc *cis*-régulés), et cette tendance diminuait avec la distance. Notons ici que l'origine de la *cis*-régulation des PCG aux alentours du LNC *cis*-régulé n'est pas forcément claire : s'agit-il d'un contrôle par le même variant *cis*-régulateur que le LNC, ou bien est-ce le LNC qui régule en *cis* ses gènes ? On voit en tous cas que la classification des LNC avec le PCG le plus proche est

pertinente, et qu'il pourrait être éventuellement intéressant d'étudier les corrélations de tous les gènes à proximité, et pas uniquement le plus proche. Notons enfin que le fait qu'un variant *cis*-régulateur puisse réguler, directement ou en régulant un LNC lui-même *cis*-régulateur, l'expression de gènes dans une région pourrait expliquer les observations que nous avons faites de régions contenant des gènes co-localisés et co-exprimés dans la partie II (article 4).

Un éventuel inconvénient de notre annotation qui se base sur les couples LNC:PCG vient de l'incertitude de la modélisation des LNC. En effet, il est possible qu'un LNC ne soit en réalité qu'un morceau d'un PCG proche, que l'outil de modélisation n'a pas réussi à relier, ou bien même que plusieurs LNC proches n'en fassent en fait qu'un. De tels phénomènes risquent de biaiser les statistiques de co-expression. Nous l'avons constaté avec les classes « même-brin » ou « sens d'intron », qui sont généralement plus enrichies en couples LNC:PCG co-exprimés que les autres. Notons que si ces classes représentent respectivement 44% et 11% des couples LNC:PCG, elles représentent également 40% et 6%, respectivement des couples PCG:PCG, soit des proportions sensiblement similaires. Le fait que plusieurs LNC en même-brin puissent en fait n'être qu'un seul LNC peut également induire un biais, car chacun d'entre eux pourrait être classifié de la même manière par rapport à un même PCG. Si tous sont co-exprimés avec lui, ils risquent d'enfler artificiellement le nombre de couples co-exprimés dans la classification à laquelle ils appartiennent. En première approche, on pourrait considérer que pour s'affranchir de ce problème, tous les LNC ayant ne serait-ce qu'un transcrit classifié comme « même-brin » avec un PCG pourraient être considérés comme suspect. Cette approche est néanmoins très stricte. En effet, en plus de la proportion similaire de ces couples parmi les couples PCG:PCG, rappelons que nous avons mis en évidence dans les articles 1 et 3 (dans la partie consacrée aux LNC, § *g*) des PCG co-localisés en même-brin et co-exprimés. Cela laisse à penser que le critère « même-brin et co-exprimé » n'est pas nécessairement indicateur d'une erreur. Pour remédier à ce problème, il risque d'être nécessaire de procéder à des validations de l'existence ou non d'un seul modèle réunissant les gènes entre eux, après avoir détecté les cas suspects en croisant (*i*) l'annotation et (*ii*) des analyses de co-expression au sein de différents tissus. Le problème ici est bien sûr le coût en travail nécessaire à une telle validation, dès lors que l'on dépasse la dizaine de gènes. Une solution intermédiaire pourrait être d'analyser, pour les cas suspects, le nombre de *reads* de RNA-seq chevauchant les deux modèles en utilisant d'autres sources que celles qui ont servies à la modélisation des gènes. C'est en effet en observant les *reads* de RNA-seq sur IGV et la manière dont ils joignent, ou non, les modèles entre eux que l'on acquiert la quasi-certitude de l'existence propre d'un LNC ou non, et dans ce dernier cas, que l'on est

capable de suggérer le PCG auquel il appartient. En effet, le LNC ne fait pas forcément partie du PCG le plus proche. Ensuite, nous avons observé dans l'article 3 (dans la partie consacrée aux LNC, § *g*) que, dans les couples LNC:PCG, lorsqu'un des membres était différentiellement exprimé, l'autre membre l'était également plus souvent que dans les couples PCG:PCG : en moyenne, cela concernait 35% des couples LNC:PCG avec au moins un membre DE contre environ 15% pour les PCG:PCG. Cela suggère que les couples LNC:PCG ont une tendance plus forte que les couples PCG:PCG à être co-régulés, et donc à être impliqués dans des processus biologiques similaires. Il serait dès lors intéressant d'étudier les mécanismes qui sont à l'origine de cette co-expression, et les rôles que jouent le PCG et le LNC. En ce qui concerne les mécanismes à l'origine de la co-expression, on peut penser que le LNC agit sur l'expression du PCG, ou bien qu'ils sont tous les deux régulés par un même mécanisme (un promoteur commun bidirectionnel dans le cas des couples divergents). Pour ce qui est des rôles respectifs de chaque gène, le LNC peut, on l'a dit, n'être « que » régulateur de l'expression du PCG, ou bien jouer un rôle dans la même voie métabolique que le PCG et jouer un autre rôle que régulateur de son PCG « compagnon ».

Il nous semble ainsi que l'approche de classification que nous avons adoptée, les informations que nous fournissons dans notre annotation étendue (notamment de corrélation d'expression et d'expression tissue-spécifique), et les différents commentaires que nous faisons des situations « suspectes » fournissent une ressource précieuse et sûre pour aider à une meilleure annotation fonctionnelle des LNC.

III – Variants génomiques : la *terra incognita* du génome

Notre travail sur l'étude de variants de type SNP par RNA-seq ouvre pour sa part différentes opportunités, notamment pour l'annotation fonctionnelle des variants affectant les régions transcrites, codantes ou non, à travers la ré-analyse des jeux de données RNA-seq qui se sont accumulés ces dernières années. Ces variants sont en effet intéressants à deux titres : ils peuvent être causaux d'un caractère complexe ou permettre d'inférer la fonction d'un gène.

En ce qui concerne les premiers, même si on s'attend à ce que les variants causaux d'un caractère complexe soient majoritairement localisés dans les régions régulatrices de l'expression, comme nous l'avons vu à différentes reprises au cours de la présente thèse, la majeure partie des variants causaux identifiés à ce jour sont localisés dans les régions codantes. Cela vient d'abord de ce que les outils de prédiction des conséquences et de la gravité des variants permettent de mettre rapidement en évidence des variants de ce type dans un jeu de données. En effet, la connaissance du code génétique permet de prédire l'effet d'un variant génomique sur la séquence d'acides-aminés de la protéine, et des études de conservation des acides-aminés à travers les espèces[391,392] ou bien des hypothèses biologiques (on peut par exemple penser que l'apparition d'un codon stop très tôt dans la séquence est plus grave que son apparition vers la toute fin) permettent d'en évaluer la gravité. De plus, les parties codantes du génome sont moins coûteuses à observer qu'un génome entier, puisqu'elles n'en représentent qu'une fraction : nous avons observé dans l'article 2 que les exons représentent environ 7% du génome de la poule, dont les CDS constituent les trois-quarts. Cela explique bien pourquoi le consortium Genome Aggregation Database (gnomAD)[345] utilise pour ses analyses environ 100 000 exomes (issus de *Whole Exome Sequencing*, WES) contre « seulement » 10 000 génomes entiers humains, et montre le potentiel du RNA-seq pour le même type d'études dans les espèces dans lesquelles le WES ne peut être réalisé, les espèces d'élevages en particulier. Enfin, une fois séquencées les parties codantes du génome et prédit l'impact des variants qu'elles contiennent, on observe généralement assez peu de variants délétères pour la fonction de la protéine, ce qui permet de concentrer les efforts pour éventuellement confirmer leur rôle.

En ce qui concerne l'inférence de la fonction d'un gène, on peut considérer que la présence chez un individu d'un variant codant affectant fortement la séquence de la protéine codée fournit d'une certaine façon un modèle de *knock-down* du gène en question[393]. En particulier, le consortium gnomAD s'est intéressé aux variants de type stop prématuré, décalage du cadre de lecture ou altération de sites d'épissage[345]. Nous parlerons de ces variants comme

étant des « variants délétères » dans la suite. En allant plus loin, on peut également considérer que l'absence de variant délétère dans des gènes en particulier au sein d'une espèce suggère un rôle clef pour ces gènes[345]. Ce sont là deux approches adoptées par le consortium gnomAD[345]. Quoi qu'il en soit, l'étude approfondie des effets des variants délétères dans une espèce, notamment pour étudier la fonction des gènes affectés, requiert un volume de données génomiques encore bien supérieur à celui utilisé par gnomAD (pourtant de l'ordre de grandeur de la centaine de milliers), mais également de données de phénotypage fines des individus pour lesquels les données génomiques seront disponibles[345].

La mise en évidence de variants d'intérêt par prédiction de leurs conséquences codantes reste un défi lorsque l'on s'intéresse aux variants autre que les plus délétères. D'un point de vue biologique d'abord, vu l'écrasante majorité de variants dont les impacts sont difficiles à prédire, en particulier les *missense* (37% des variants dans les régions codantes) et surtout les variants *synonymous* (62% des variants dans les régions codantes). D'un point de vue technique ensuite, les outils de prédiction des effets procèdent parfois à des simplifications qui peuvent empêcher la mise en évidence de variants d'intérêt. Ainsi, nous avons observé que ces outils ont tendance à étudier les variants les uns après les autres, en considérant que chaque variant est entouré par les nucléotides du génome de référence. Ainsi, si deux ou trois SNP sont présents dans un même triplet de nucléotides, une prédiction sera réalisée pour chaque SNP, considéré isolément, avec le risque que ces prédictions soient fausses si les SNP sont tous portés par le même chromosome. Considérons l'exemple suivant : soit le codon de départ CGU (codant pour l'arginine). Chez un individu, un polymorphisme C > U affecte le premier nucléotide et un autre polymorphisme U > A affecte le troisième nucléotide. Le codon porté par l'individu est donc UGA. Les prédictions indépendantes des conséquences fourniront pour résultats : (*i*) CGU > UGU entraîne la traduction d'une sérine, il s'agit d'un polymorphisme « faux-sens » (*missense*), (*ii*) CGU > CGA entraîne la traduction d'une l'arginine, il s'agit d'un polymorphisme « même-sens » (*synonymous*). À moins que l'arginine codée par ce codon soit particulièrement conservée entre espèces, aucun outil de prédiction ne fera ressortir le *missense*, et encore moins le *synonymous*. Pourtant, le codon UGA, qui est porté par cet individu, entraîne l'apparition d'un stop, potentiellement délétère. L'inverse est aussi vrai, et un SNP prédit comme provoquant un stop prématuré peut être accompagné d'un autre SNP, pour provoquer finalement un *missense* ou un *synonymous*. Nous avons, dans le cadre de l'encadrement d'un étudiant de master 2, participé au développement d'un script *ad hoc* utilisant les informations de phase et les conséquences prédites pour détecter ces cas et re-prédire les conséquences en tenant compte de l'existence de deux ou trois SNP dans un codon. Par simplification, nous

avons associé à ces codons la conséquence prédite la plus grave, selon l'ordre : stop prématuré > *missense* > *synonymous*. Notons que nous n'avons pas utilisé phASER pour prédire ces phases, mais GATK. Elles correspondent donc aux SNP présents sur un même *read*. Nos résultats préliminaires, utilisant les travaux de l'article 2, montrent qu'environ 0.2% (18 356 SNP) des ~10 millions de SNP détectés étaient phasés et présents par 2 dans le même codon, pour un total de 9 178 codons. Parmi ces 9 178 codons, 93% (8 581 codons) étaient prédits comme étant des *missense*, et parmi eux, 2% (177 codons) ont été reprédits comme stop. À l'inverse, 3.8% des codons (346 codons) étaient prédits comme devenant un stop, et parmi eux, 333 (96% des stop prédits) étaient en fait des *missense*. Des résultats similaires ont été obtenus pour les 246 codons présentant 3 SNP phasés : 84% des 25 prédictions de stop prématurés étaient des *missense* (21 codons), et 6% des *missense* étaient en fait des stop prématurés (13 codons).

L'étude des SNP par RNA-seq et la prédiction de leurs conséquences représentent donc une perspective intéressante pour l'annotation des gènes codants des protéines, en particulier chez les espèces d'élevages chez lesquelles les moyens techniques et financiers ne permettent pas de produire autant de données de séquençage d'exomes (WES, *Whole Exome Sequencing*) et de génomes entiers (WGS, *Whole Genome Sequencing*).

Reste enfin le défi de la prédiction des conséquences des variants affectant les LNC. Nous avons vu que les LNC étaient moins bien conservés en séquence entre espèces que les PCG, et que leurs séquences primaires subissaient une pression de sélection moins forte que celles des PCG. En revanche, l'importance apparente de la composition en *k*-mers suggère que des portions de séquence sont importants à la fonction du LNC, mais en l'absence de régions fonctionnelles bien définies et de mécanismes d'action connus, il est encore impossible de prédire une fonction à partir des *k*-mer et ainsi de déterminer si un SNP serait à même d'avoir un effet ou non, sauf si ce sont des variants plus grands, comme les insertions-délétions, qui affectent plutôt les LNC. Nous avons pris ce parti dans l'article 5 (en préparation) en étudiant les insertions-délétions (INDEL) affectant les LNC et dont les fréquences étaient très contrastées entre les deux lignées et dont la taille était supérieure ou égale à celle des *k*-mers de Kirk *et al*.[258] (2018), soit $k \geq 6$. Cette approche nous a permis de mettre en évidence 4 LNC qu'il s'agira éventuellement d'explorer plus en détail, notamment via la corrélation d'expression avec les PCG proches à travers de multiples tissus et individus.

En conclusion de cette thèse, nous pensons pouvoir dire que les travaux que nous avons entrepris ont permis d'un petit peu mieux comprendre les processus œuvrant aux nombreuses différences entre les lignées R+ et R-, aussi bien au niveau des tissus que des gènes potentiellement causaux. Le gène *FABP4* en particulier nous parait être un candidat tout-à-fait intéressant, et il conviendra d'essayer de valider son implication et de comprendre sa régulation. Nous avons également contribué à l'amélioration de l'annotation du génome de la poule en mettant en évidence de nouveaux LNC qui pourront être utilisés par la communauté scientifique en lien avec différentes problématiques, comme nous l'avons fait pour notre travail sur l'efficience alimentaire et les caractères associés. La détection de SNP par RNA-seq et la mise en place du pipeline ASE devraient quant à eux ouvrir la voie à des études plus poussées sur les *cis*-régulations chez la poule et dans les espèces d'élevage en lien ou non avec la détection de gène causaux de la variation d'un caractère.

Ces différents travaux nous ont permis de parfois percevoir la nécessité d'une clarification de certains concepts : ainsi par exemple, rappelons que nombre d'ARN « long non-codants » codent de petits peptides dont les rôles restent à définir. Au-delà de l'aspect contradictoire entre le nom et les fonctions ou les potentiels, cet aspect sémantique nous semble important car il guide d'une manière plus ou moins forte les recherches. En effet, on aura à priori moins tendance à vouloir prédire la conséquence codante d'un variant dans un LNC. En ce qui concerne le terme « ARN long non-codant » lui-même, il est probablement amené à être remplacé par des termes plus précis, qui ne qualifieront plus seulement ces gènes par ce qu'ils ne font pas (à savoir coder des protéines, ce qui est plutôt erroné comme nous venons de le voir), mais bien pour ce qu'ils font, lorsque suffisamment de travaux seront parvenus à élucider leurs rôles.

Pour ce qui est des gènes *cis*-régulés, nous avons pu constater qu'il est relativement facile de les mettre en évidence avec des données pertinentes, mais à quel point il est compliqué de comprendre leurs rôles. À cet égard, on ne peut qu'espérer que le consortium gnomAD, par son « retour » aux variants dans les régions codantes pourra fournir des annotations fonctionnelles plus vastes et plus précises, au moins pour les PCG.

Il reste, on le voit, bien du travail à accomplir pour un jour pouvoir embrasser dans leur ensemble les différents mécanismes biologiques, interactions et régulations associés à un caractère complexe, et leur fascinant ballet.

Bibliographie

1. *Livestock in the balance*. (FAO, 2009).
2. *Climate change 2007: mitigation of climate change: contribution of Working Group III to the Fourth assessment report of the Intergovernmental Panel on Climate Change*. (Cambridge University Press, 2007).
3. *Tackling climate change through livestock: a global assessment of emissions and mitigation opportunities*. (Food and Agriculture Organization of the United Nations, 2013).
4. Rojas-Downing, M. M., Nejadhashemi, A. P., Harrigan, T. & Woznicki, S. A. Climate change and livestock: Impacts, adaptation, and mitigation. *Climate Risk Management* **16**, 145–163 (2017).
5. Présentation | Présentation | Organisation des filières : les différents acteurs | Contenu du cours 40003 | FUN-MOOC. https://www.fun-mooc.fr/courses/course-v1:agrocampusouest+40003+session02/courseware/dce3e1d1ac184000a4a973e79f7b3a13/776d33a8be154934bc6b1dcc8bb03ed4/.
6. Food and Agriculture Organization of the United Nations. *Livestock & Climate Change*. 16 http://www.fao.org/3/a-i6345e.pdf (2016).
7. McGrath, J. *et al.* Nutritional strategies in ruminants: A lifetime approach. *Research in Veterinary Science* **116**, 28–39 (2018).
8. Patience, J. F., Rossoni-Serão, M. C. & Gutiérrez, N. A. A review of feed efficiency in swine: biology and application. *J Animal Sci Biotechnol* **6**, 33 (2015).
9. Wen, C. *et al.* Feed efficiency measures and their relationships with production and meat quality traits in slower growing broilers. *Poultry Science* **97**, 2356–2364 (2018).
10. Gabarrou, J. F. *et al.* Energy balance of laying hens selected on residual food consumption. *British Poultry Science* **39**, 79–89 (1998).
11. Ryschawy, J. *et al.* Assessing multiple goods and services derived from livestock farming on a nation-wide gradient. *Animal* **11**, 1861–1872 (2017).
12. Ryschawy, J. *et al.* Comment évaluer les services rendus par l'élevage ? Une première approche méthodologique sur le cas de la France. 16 (2015).
13. Dumont, B. *et al. Rôles, impacts et services issus des élevages en Europe*. 1032 (2016).
14. Leitch, I. & Godden, W. Efficiency of Farm Animals in Feed Conversion. *Can J Comp Med Vet Sci* **5**, 292–293 (1941).
15. Byerly, T. C. & others. Feed and other costs of producing market eggs. (1941).
16. Kleiber, M. Problems involved in breeding for efficiency of food utilization. *Journal of Animal Science* **1936b**, 247–258 (1936).
17. Hess, C. W. & Jull, M. A. A Study of the Inheritance of Feed Utilization Efficiency in the Growing Domestic Fowl. *Poultry Science* **27**, 24–39 (1948).
18. Dickerson, G. E. & Gowen, J. W. Hereditary Obesity and Efficient Food Utilization in Mice. *Science* **105**, 496–498 (1947).
19. Morris, H. P., Palmer, L. S. & Kennedy, C. Fundamental food requirements for the growth of the rat: VII, An experimental study of inheritance as a factor influencing food utilization in the rat. *Univ. Minnesota, Agric. Exp. Stat.* 56 (1933).
20. Titus, H. W., Mehring, A. L. & Brumbaugh, J. H. Variation of Feed Conversion. *Poultry Science* **32**, 1074–1077 (1953).

21. Koch, R. M., Swiger, L. A., Chambers, D. & Gregory, K. E. Efficiency of Feed Use in Beef Cattle. *Journal of Animal Science* **22**, 486–494 (1963).
22. Taking control of feed conversion ratio. *PigProgress* https://www.pigprogress.net/Breeding/Sow-Feeding/2009/4/Taking-control-of-feed-conversion-ratio-PP005927W/.
23. Ceballos, L. S. *et al.* Composition of goat and cow milk produced under similar conditions and analyzed by identical methodology. *Journal of Food Composition and Analysis* **22**, 322–329 (2009).
24. Food and Agriculture Organization of the United Nations. *The state of the world fisheries and aquaculture 2014: opportunities and challenges.* (Food and Agriculture Organization of the United Nations, 2014).
25. IPIFF. *The european insect sector today: challenges, opportunities and regulatory landscape*. (2019).
26. Food and Agriculture Organization of the United Nations & Animal Production and Health Division. *Greenhouse gas emissions from pig and chicken supply chains: a global life cycle assessment*. (2013).
27. Novogen - Produits. https://www.novogen-layer.com/fr/produits/novogen-white-light/7766-commerciale-novogen-white-light.html.
28. Wilkinson, J. M. Re-defining efficiency of feed use by livestock. *Animal* **5**, 1014–1022 (2011).
29. Shike, D. W. Beef Cattle Feed Efficiency. *Driftless Region Beef Conference 2013* 2 (2013).
30. Chambre d'agriculture d'Ille-et-Vilaine. *Synthèse : résultats technico-économiques, enquête réalisée auprès des aviculteurs du Grand-Ouest*. (2014).
31. Accueil - Hybrid. https://www.hybridturkeys.com/fr/.
32. Pig production, Managing a piggery : Performance standards. https://www.daf.qld.gov.au/business-priorities/agriculture/animals/pigs/piggery-management/production-performance/standards (2013).
33. Feed Efficiency and Its Impact on Feed Intake – DAIReXNET. https://dairy-cattle.extension.org/feed-efficiency-and-its-impact-on-feed-intake/.
34. Réseaux d'élevage bovins viande. *Produire des jeunes bovins dans l'Est : 4 itinéraires techniques*. (2008).
35. Knott, S., Leury, B., Cummins, L., Brien, F. & Dunshea, F. Relationship between body composition, net feed intake and gross feed conversion efficiency in composite sire line sheep. *Progress in Research on Energy and Protein Metabolism* 525–528 (2003).
36. Malik, R. C., Razzaque, M. A., Abbas, S., Al-Khozam, N. & Sahni, S. Feedlot growth and efficiency of three-way cross lambs as affected by genotype, age and diet. *Proc. Aust. Soc. Anim. Prod.* **21**, 251–254 (1996).
37. DeLong, D. P., Losordo, T. M. & Rakocy, J. E. Tank Culture of Tilapia. *SRAC Publication* 8 (2009).
38. FAO Fisheries & Aquaculture - Cultured Aquatic Species Information Programme - Salmo salar (Linnaeus, 1758). http://www.fao.org/fishery/culturedspecies/Salmo_salar/en.

39. van Huis, A. Potential of Insects as Food and Feed in Assuring Food Security. *Annu. Rev. Entomol.* **58**, 563–583 (2013).
40. Alexander, P. *et al.* Could consumption of insects, cultured meat or imitation meat reduce global agricultural land use? *Global Food Security* **15**, 22–32 (2017).
41. FAOSTAT, Livestock Primary. http://www.fao.org/faostat/en/#data/QL.
42. Herry, F. Stratégies de génotypage pour la sélection génomique chez la poule pondeuse. 283.
43. Tixier-Boichard, M., Bed'hom, B. & Rognon, X. Chicken domestication: From archeology to genomics. *Comptes Rendus Biologies* **334**, 197–204 (2011).
44. Miao, Y.-W. *et al.* Chicken domestication: an updated perspective based on mitochondrial genomes. *Heredity* **110**, 277–282 (2013).
45. Wang, Y. Aperçu historique de l'élevage des poulets. *jatba* **28**, 253–258 (1981).
46. Wood-Gush, D. G. M. A History of the Domestic Chicken from Antiquity to the 19th Century. *Poultry Science* **38**, 321–326 (1959).
47. César, J. *Commentarii de Bello Gallico (La guerre des Gaules)*. (Les Belles Lettres, 1926).
48. Columelle : livre 8 : traduction française. http://remacle.org/bloodwolf/erudits/columelle/livre8fr.htm.
49. Elson, H. A., Gleadthorpe, A., Vale, M. & Uk, M. Housing and Husbandry of Laying Hens: past, present and future. *Lohmann information* **46**, 16 (2011).
50. Le Bohec, Y. *La guerre romaine: 58 avant J.-C. - 235 après J.-C*. (Tallandier, 2017).
51. Sheldon, R. M. *Renseignement et espionnage dans la Rome antique*. (Les Belles lettres, 2009).
52. MacDonald, K. C. The Domestic Chicken (Gallus gallus) in Sub-Saharan Africa: A Background to its Introduction and its Osteological Differentiation from Indigenous Fowls (Numidinae and Francolinus sp.). *Journal of Archaeological Science* **19**, 303–318 (1992).
53. Mwacharo, J. M., Bjørnstad, G., Han, J. L. & Hanotte, O. The History of African Village Chickens: an Archaeological and Molecular Perspective. *Afr Archaeol Rev* **30**, 97–114 (2013).
54. Fitzpatrick, S. M. & Callaghan, R. Examining dispersal mechanisms for the translocation of chicken (Gallus gallus) from Polynesia to South America. *Journal of Archaeological Science* **36**, 214–223 (2009).
55. Rubin, C.-J. *et al.* Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**, 587–591 (2010).
56. Karlsson, A.-C. *et al.* The Effect of a Mutation in the Thyroid Stimulating Hormone Receptor (TSHR) on Development, Behaviour and TH Levels in Domesticated Chickens. *PLoS ONE* **10**, e0129040 (2015).
57. Girdland Flink, L. *et al.* Establishing the validity of domestication genes using DNA from ancient chickens. *Proceedings of the National Academy of Sciences* **111**, 6184–6189 (2014).
58. Grottesi, A., Gabbianelli, F., Valentini, A. & Chillemi, G. Structural and dynamic analysis of G558R mutation in chicken *TSHR* gene shows altered signal transduction and corroborates its role as a domestication gene. *Anim Genet* **51**, 51–57 (2020).
59. Auteur Inconnu. *Le ménagier de Paris, traité de morale et d'économie domestique composé vers 1393 par un bourgeois parisien*. (imp. de Crapelet, 1846).

60. Le coq. *elysee.fr* https://www.elysee.fr/la-presidence/le-coq.
61. Bennett, C. E. *et al.* The broiler chicken as a signal of a human reconfigured biosphere. *R. Soc. open sci.* **5**, 180325 (2018).
62. FAOSTAT, Value of Agricultural Production. http://www.fao.org/faostat/en/#data/QV.
63. CNPO - Interprofession des Oeufs. Infos filière : les chiffres clés. *CNPO Site filière* https://oeuf-info.fr/infos-filiere/les-chiffres-cles/.
64. Alexandratos, N. World Agriculture towards 2030/2050: the 2012 revision. 154.
65. Village chicken production systems in rural Africa Household food security and gender issues. http://www.fao.org/3/W8989E/W8989E00.htm.
66. Melesse, A. Significance of scavenging chicken production in the rural community of Africa for enhanced food security. *World's Poultry Science Journal* **70**, 593–606 (2014).
67. Mapiye, C. *et al.* A Research Review of Village Chicken Production Constraints and Opportunities in Zimbabwe. *Asian Australas. J. Anim. Sci* **21**, 1680–1688 (2008).
68. Iannotti, L. L., Lutter, C. K., Bunn, D. A. & Stewart, C. P. Eggs: the uncracked potential for improving maternal and young child nutrition among the world's poor. *Nutr Rev* **72**, 355–368 (2014).
69. Lesnierowski, G. & Stangierski, J. What's new in chicken egg research and technology for human health promotion? - A review. *Trends in Food Science & Technology* **71**, 46–51 (2018).
70. Kaspers, B. An egg a day - the physiology of egg formation. *Lohmann information* **50**, 12–17 (2016).
71. Sauveur, B. *Reproduction des volailles et production d'œufs*. (INRA, 1988).
72. Nys, Y. & Guyot, N. Egg formation and chemistry. in *Improving the Safety and Quality of Eggs and Egg Products* 83–132 (Elsevier, 2011). doi:10.1533/9780857093912.2.83.
73. Guyot, N. *et al.* Characterization of egg white antibacterial properties during the first half of incubation: A comparative study between embryonated and unfertilized eggs. *Poultry Science* **95**, 2956–2970 (2016).
74. *Bioactive egg compounds*. (Springer, 2007).
75. Sorensen, R. A. The Egg Came Before the Chicken. *Mind* **101**, 541–542 (1992).
76. Mozdziak, P. E. & Petitte, J. N. Status of transgenic chicken models for developmental biology. *Dev. Dyn.* **229**, 414–421 (2004).
77. Dodgson, J. B. & Romanov, M. N. Use of Chicken Models for the Analysis of Human Disease. *Current Protocols in Human Genetics* **40**, (2004).
78. Glick, B., Chang, T. S. & Jaap, R. G. The Bursa of Fabricius and Antibody Production. *Poultry Science* **35**, 224–225 (1956).
79. Kumar, S. & Hedges, S. B. A molecular timescale for vertebrate evolution. *Nature* **392**, 917–920 (1998).
80. Masabanda, J. S. *et al.* Molecular Cytogenetic Definition of the Chicken Genome: The First Complete Avian Karyotype. *Genetics* **166**, 1367–1373 (2004).
81. Cheng, Y. & Burt, D. W. Chicken genomics. *Int. J. Dev. Biol.* **62**, 265–271 (2018).
82. Solinhac, R. *et al.* Integrative mapping analysis of chicken microchromosome 16 organization. *BMC Genomics* **11**, 616 (2010).
83. Warren, W. C. *et al.* A New Chicken Genome Assembly Provides Insight into Avian Genome Structure. *G3* **7**, 109–117 (2017).

84. King, M. & Wilson, A. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).
85. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. **322**, 5 (2008).
86. Fisher, R. A. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Trans. R. Soc. Edinb.* **52**, 399–433 (1919).
87. Risch, N. *et al.* A Genomic Screen of Autism: Evidence for a Multilocus Etiology. *The American Journal of Human Genetics* **65**, 493–507 (1999).
88. The Electronic Medical Records and Genomics (eMERGE) Consortium *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* **46**, 1173–1186 (2014).
89. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
90. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565–569 (2010).
91. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. *The American Journal of Human Genetics* **99**, 139–153 (2016).
92. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017).
93. Pickrell, J. K. Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits. *The American Journal of Human Genetics* **94**, 559–573 (2014).
94. Li, Y. I. *et al.* RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600–604 (2016).
95. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* 1222794 (2012).
96. Gross, D. S. & Garrard, W. T. Nuclease Hypersensitive Sites in Chromatin. *Annu. Rev. Biochem.* **57**, 159–197 (1988).
97. Boyle, A. P. *et al.* High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell* **132**, 311–322 (2008).
98. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580–585 (2013).
99. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
100. Dixon, J. R., Gorkin, D. U. & Ren, B. Chromatin Domains: The Unit of Chromosome Organization. *Molecular Cell* **62**, 668–680 (2016).
101. Szabo, Q., Bantignies, F. & Cavalli, G. Principles of genome folding into topologically associating domains. *Sci. Adv.* **5**, eaaw1668 (2019).
102. Foissac, S. *et al.* Multi-species annotation of transcriptome and chromatin structure in domesticated animals. *BMC Biol* **17**, 108 (2019).
103. Fishman, V. *et al.* 3D organization of chicken genome demonstrates evolutionary conservation of topologically associated domains and highlights unique architecture of erythrocytes' chromatin. 18.

104. Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nat Rev Genet* **7**, 85–97 (2006).
105. The ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
106. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
107. Shendure, J., Findlay, G. M. & Snyder, M. W. Genomic Medicine–Progress, Pitfalls, and Promise. *Cell* **177**, 45–57 (2019).
108. Le Bihan-Duval, E. *et al.* Detection of a Cis eQTL Controlling BMCO1 Gene Expression Leads to the Identification of a QTG for Chicken Breast Meat Color. *PLoS ONE* **6**, e14825 (2011).
109. Plassais, J. *et al.* A Point Mutation in a lincRNA Upstream of GDNF Is Associated to a Canine Insensitivity to Pain: A Spontaneous Model for Human Sensory Neuropathies. *PLoS Genet* **12**, e1006482 (2016).
110. Tian, J. *et al.* Systematic Functional Interrogation of Genes in GWAS Loci Identified ATF1 as a Key Driver in Colorectal Cancer Modulated by a Promoter-Enhancer Interaction. *The American Journal of Human Genetics* **105**, 29–47 (2019).
111. Lagarrigue, S. *et al.* Analysis of Allele-Specific Expression in Mouse Liver by RNA-Seq: A Comparison With *Cis* -eQTL Identified Using Genetic Linkage. *Genetics* **195**, 1157–1166 (2013).
112. van Nas, A. *et al.* Expression Quantitative Trait Loci: Replication, Tissue- and Sex-Specificity in Mice. *Genetics* **185**, 1059–1068 (2010).
113. Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics* **16**, 197–212 (2015).
114. Barr, M. L. & Bertram, E. G. A Morphological Distinction between Neurones of the Male and Female, and the Behaviour of the Nucleolar Satellite during Accelerated Nucleoprotein Synthesis. *Nature* **163**, 676–677 (1949).
115. DeChiara, T. M., Robertson, E. J. & Efstratiadis, A. Parental imprinting of the mouse insulin-like growth factor II gene. *Cell* **64**, 849–859 (1991).
116. Surani, M. A. H., Barton, S. C. & Norris, M. L. Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. *Nature* **308**, 548–550 (1984).
117. Giannoukakis, N., Deal, C., Paquette, J., Goodyer, C. G. & Polychronakos, C. Parental genomic imprinting of the human IGF2 gene. *Nat Genet* **4**, 98–101 (1993).
118. Baran, Y. *et al.* The landscape of genomic imprinting across diverse adult human tissues. *Genome Res.* **25**, 927–936 (2015).
119. Prickett, A. R. & Oakey, R. J. A survey of tissue-specific genomic imprinting in mammals. *Mol Genet Genomics* **287**, 621–630 (2012).
120. Gimelbrant, A., Hutchinson, J. N., Thompson, B. R. & Chess, A. Widespread Monoallelic Expression on Human Autosomes. *Science* **318**, 1136–1140 (2007).
121. Zwemer, L. M. *et al.* Autosomal monoallelic expression in the mouse. *Genome Biol* **13**, R10 (2012).
122. Reinius, B. *et al.* Analysis of allelic expression patterns in clonal somatic cells by single-cell RNA–seq. *Nat Genet* **48**, 1430–1435 (2016).

123. Frésard, L. *et al.* Transcriptome-wide investigation of genomic imprinting in chicken. *Nucleic Acids Research* **42**, 3768–3782 (2014).
124. Wang, Q. *et al.* Next-Generation Sequencing Techniques Reveal that Genomic Imprinting Is Absent in Day-Old Gallus gallus domesticus Brains. *PLoS ONE* **10**, e0132345 (2015).
125. Zhuo, Z., Lamont, S. J. & Abasht, B. RNA-Seq Analyses Identify Frequent Allele Specific Expression and No Evidence of Genomic Imprinting in Specific Embryonic Tissues of Chicken. *Sci Rep* **7**, 11944 (2017).
126. Wang, Q., Mank, J. E., Li, J., Yang, N. & Qu, L. Allele-Specific Expression Analysis Does Not Support Sex Chromosome Inactivation on the Chicken Z Chromosome. *Genome Biology and Evolution* **9**, 619–626 (2017).
127. Knight, J. C. Allele-specific gene expression uncovered. *Trends in Genetics* **20**, 113–116 (2004).
128. Buckland, P. R. Allele-specific gene expression differences in humans. *Human Molecular Genetics* **13**, R255–R260 (2004).
129. Chamberlain, A. J. *et al.* Extensive variation between tissues in allele specific expression in an outbred mammal. *BMC Genomics* **16**, 993 (2015).
130. Guillocheau, G. M. *et al.* Survey of allele specific expression in bovine muscle. *Sci Rep* **9**, 4297 (2019).
131. Ren, P. *et al.* Genome-wide analysis of spatiotemporal allele-specific expression in F1 hybrids of meat- and egg-type chickens. *Gene* **747**, 144671 (2020).
132. Maroilley, T. *et al.* Deciphering the genetic regulation of peripheral blood transcriptome in pigs through expression genome-wide association study and allele-specific expression analysis. *BMC Genomics* **18**, 967 (2017).
133. Edsgärd, D. *et al.* GeneiASE: Detection of condition-dependent and static allele-specific expression from RNA-seq data without haplotype information. *Sci Rep* **6**, 21134 (2016).
134. Fan, J. *et al.* ASEP: Gene-based detection of allele-specific expression across individuals in a population by RNA sequencing. *PLoS Genet* **16**, e1008786 (2020).
135. Jastrebski, S. F., Lamont, S. J. & Schmidt, C. J. Chicken hepatic response to chronic heat stress using integrated transcriptome and metabolome analysis. *PLoS ONE* **12**, e0181900 (2017).
136. Srikanth, K. *et al.* Genome-Wide Transcriptome and Metabolome Analyses Provide Novel Insights and Suggest a Sex-Specific Response to Heat Stress in Pigs. *Genes* **11**, 540 (2020).
137. Videvall, E., Cornwallis, C. K., Palinauskas, V., Valkiūnas, G. & Hellgren, O. The Avian Transcriptome Response to Malaria Infection. *Molecular Biology and Evolution* **32**, 1255–1267 (2015).
138. Skugor, A. *et al.* Effects of long-term feeding of rapeseed meal on skeletal muscle transcriptome, production efficiency and meat quality traits in Norwegian Landrace growing-finishing pigs. *PLoS ONE* **14**, e0220441 (2019).
139. Evans, T. G. Considerations for the use of transcriptomics in identifying the 'genes that matter' for environmental adaptation. *Journal of Experimental Biology* **218**, 1925–1935 (2015).

140. López-Maury, L., Marguerat, S. & Bähler, J. Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nat Rev Genet* **9**, 583–593 (2008).
141. Nadeau, S. I. & Landry, J. Mechanisms of Activation and Regulation of the Heat Shock-Sensitive Signaling Pathways. in *Molecular Aspects of the Stress Response: Chaperones, Membranes and Networks* (eds. Csermely, P. & Vígh, L.) vol. 594 100–113 (Springer New York, 2007).
142. Efeyan, A., Comb, W. C. & Sabatini, D. M. Nutrient-sensing mechanisms and pathways. *Nature* **517**, 302–310 (2015).
143. Yang, T. *et al.* Crucial Step in Cholesterol Homeostasis: Sterols Promote Binding of SCAP to INSIG-1, a Membrane Protein that Facilitates Retention of SREBPs in ER. 12 (2002).
144. Radhakrishnan, A., Goldstein, J. L., McDonald, J. G. & Brown, M. S. Switch-like Control of SREBP-2 Transport Triggered by Small Changes in ER Cholesterol: A Delicate Balance. *Cell Metabolism* **8**, 512–521 (2008).
145. Motamed, M. *et al.* Identification of Luminal Loop 1 of Scap Protein as the Sterol Sensor That Maintains Cholesterol Homeostasis. *J. Biol. Chem.* **286**, 18002–18012 (2011).
146. Jeon, T.-I. & Osborne, T. F. SREBPs: metabolic integrators in physiology and metabolism. *Trends in Endocrinology & Metabolism* **23**, 65–72 (2012).
147. Lister, R. *et al.* Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell* **133**, 523–536 (2008).
148. Nagalakshmi, U. *et al.* The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science* **320**, 1344–1349 (2008).
149. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621–628 (2008).
150. Bainbridge, M. N. *et al.* Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* **7**, 246 (2006).
151. *Molecular cell biology*. (Freeman, 2002).
152. Kukurba, K. R. & Montgomery, S. B. RNA Sequencing and Analysis. *Cold Spring Harb Protoc* **2015**, pdb.top084970 (2015).
153. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17**, 333–351 (2016).
154. Top 10 Gene Sequencing Companies by Revenue. *BioSpace* https://www.biospace.com/article/top-10-gene-sequencing-companies-by-revenue/.
155. Martin, J. A. & Wang, Z. Next-generation transcriptome assembly. *Nat Rev Genet* **12**, 671–682 (2011).
156. Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M. & Kelly, S. TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome Res.* **26**, 1134–1144 (2016).
157. Geniza, M. & Jaiswal, P. Tools for building de novo transcriptome assembly. *Current Plant Biology* **11**–**12**, 41–45 (2017).
158. Zhang, G. *et al.* Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**, 1311–1320 (2014).
159. Hnilicová, J. & Staněk, D. Where splicing joins chromatin. *Nucleus* **2**, 182–188 (2011).

160. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).
161. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525–527 (2016).
162. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
163. The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636–640 (2004).
164. Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **131**, 281–285 (2012).
165. Zhao, S., Ye, Z. & Stanton, R. Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *RNA* **26**, 903–909 (2020).
166. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
167. Chen, Y., McCarthy, D., Ritchie, M., Robinson, M. & Smyth, G. edgeR: differential analysis of sequence read count data - User's Guide. 121.
168. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology* **11**, R25 (2010).
169. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
170. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* **12**, 323 (2011).
171. Van den Berge, K. *et al.* RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis. *Annu. Rev. Biomed. Data Sci.* **2**, 139–173 (2019).
172. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
173. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
174. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24–26 (2011).
175. Struhl, K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat Struct Mol Biol* **14**, 103–105 (2007).
176. Clark, M. B. *et al.* The Reality of Pervasive Transcription. *PLoS Biol* **9**, e1000625 (2011).
177. Muret, K. Annotation des ARN longs non-codants chez la poule et les espèces d'élevage: Focus sur les ARNlnc régulateurs du métabolisme des lipides. 281.
178. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511–515 (2010).
179. Behm, M. & Öhman, M. RNA Editing: A Contributor to Neuronal Dynamics in the Mammalian Brain. *Trends in Genetics* **32**, 165–175 (2016).
180. Lagarrigue, S. *et al.* Limited RNA Editing in Exons of Mouse Liver and Adipose. *Genetics* **193**, 1107–1115 (2013).

181. Porath, H. T., Carmi, S. & Levanon, E. Y. A genome-wide map of hyper-edited RNA reveals numerous new sites. *Nature Communications* **5**, 4726 (2014).
182. Olofsson, B. & Bernardi, G. The distribution of CR1, an Alu-like family of interspersed repeats, in the chicken genome. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression* **740**, 339–341 (1983).
183. Carmi, S., Borukhov, I. & Levanon, E. Y. Identification of Widespread Ultra-Edited Human RNAs. *PLoS Genet* **7**, e1002317 (2011).
184. Ramaswami, G. & Li, J. B. RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucl. Acids Res.* **42**, D109–D113 (2014).
185. Tan, M. H. Dynamic landscape and regulation of RNA editing in mammals. 27.
186. Roux, P.-F. *et al.* The Extent of mRNA Editing Is Limited in Chicken Liver and Adipose, but Impacted by Tissular Context, Genotype, Age, and Feeding as Exemplified with a Conserved Edited Site in COG3. *G3* **6**, 321–335 (2016).
187. Frésard, L. *et al.* Genome-Wide Characterization of RNA Editing in Chicken Embryos Reveals Common Features among Vertebrates. *PLoS ONE* **10**, e0126776 (2015).
188. Kleinman, C. L., Adoue, V. & Majewski, J. RNA editing of protein sequences: A rare event in human transcriptomes. *RNA* **18**, 1586–1596 (2012).
189. Li, M. *et al.* Widespread RNA and DNA Sequence Differences in the Human Transcriptome. *Science* **333**, 53–58 (2011).
190. Piskol, R., Ramaswami, G. & Li, J. B. Reliable Identification of Genomic Variants from RNA-Seq Data. *The American Journal of Human Genetics* **93**, 641–651 (2013).
191. Quinn, E. M. *et al.* Development of Strategies for SNP Detection in RNA-Seq Data: Application to Lymphoblastoid Cell Lines and Evaluation Using 1000 Genomes Data. *PLoS ONE* **8**, e58815 (2013).
192. Tang, X. *et al.* The eSNV-detect: a computational system to identify expressed single nucleotide variants from transcriptome sequencing data. *Nucleic Acids Research* **42**, e172–e172 (2014).
193. Wang, C. *et al.* RVboost: RNA-seq variants prioritization using a boosting method. *Bioinformatics* **30**, 3414–3416 (2014).
194. Wolfien, M. *et al.* TRAPLINE: a standardized and automated pipeline for RNA sequencing data analysis, evaluation and annotation. *BMC Bioinformatics* **17**, 21 (2016).
195. Guo, Y., Zhao, S., Sheng, Q., Samuels, D. C. & Shyr, Y. The discrepancy among single nucleotide variants detected by DNA and RNA high throughput sequencing data. *BMC Genomics* **18**, 690 (2017).
196. Oikkonen, L. & Lise, S. Making the most of RNA-seq: Pre-processing sequencing data with Opossum for reliable SNP variant detection. *Wellcome Open Res* **2**, 6 (2017).
197. Cornwell, M. *et al.* VIPER: Visualization Pipeline for RNA-seq, a Snakemake workflow for efficient and complete RNA-seq analysis. *BMC Bioinformatics* **19**, 135 (2018).
198. Adetunji, M. O., Lamont, S. J., Abasht, B. & Schmidt, C. J. Variant analysis pipeline for accurate detection of genomic variants from transcriptome sequencing data. *PLoS ONE* **14**, e0216838 (2019).
199. NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium *et al.* Allele-specific expression changes dynamically during T cell activation in HLA and other autoimmune loci. *Nat Genet* **52**, 247–253 (2020).

200. Rao, X. *et al.* Allele-specific expression and high-throughput reporter assay reveal functional genetic variants associated with alcohol use disorders. *Mol Psychiatry* (2019) doi:10.1038/s41380-019-0508-z.
201. Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. Tools and best practices for data processing in allelic expression analysis. *Genome Biol* **16**, 195 (2015).
202. Castel, S. E., Mohammadi, P., Chung, W. K., Shen, Y. & Lappalainen, T. Rare variant phasing and haplotypic expression from RNA sequencing with phASER. *Nat Commun* **7**, 12817 (2016).
203. Castel, S. *secastel/phaser*. (2020).
204. Larson, N. B. *et al.* Comprehensively Evaluating cis -Regulatory Variation in the Human Prostate Transcriptome by Using Gene-Level Allele-Specific Expression. *The American Journal of Human Genetics* **96**, 869–882 (2015).
205. Jia, X. *et al.* Common risk variants in NPHS1 and TNFSF15 are associated with childhood steroid-sensitive nephrotic syndrome. *Kidney International* S0085253820306396 (2020) doi:10.1016/j.kint.2020.05.029.
206. GTEx Consortium *et al.* A vast resource of allelic expression data spanning human tissues. *Genome Biol* **21**, 234 (2020).
207. Oliva, M. *et al.* The impact of sex on gene expression across human tissues. *Science* **369**, eaba3066 (2020).
208. Shaffer, S. M. *et al.* Memory Sequencing Reveals Heritable Single-Cell Gene Expression Programs Associated with Distinct Cellular Behaviors. *Cell* **182**, 947-959.e17 (2020).
209. Crick, F. H. On protein synthesis. *Symp. Soc. Exp. Biol.* **12**, 138–163 (1958).
210. dogme - Définitions, synonymes, conjugaison, exemples | Dico en ligne Le Robert. https://dictionnaire.lerobert.com/definition/dogme.
211. Crick, F. *What mad pursuit: a personal view of scientific discovery*. (Basic Books, 1988).
212. Cobb, M. 60 years ago, Francis Crick changed the logic of biology. *PLoS Biol* **15**, e2003243 (2017).
213. Brenner, S., Jacob, F. & Meselson, M. An Unstable Intermediate Carrying Information from Genes to Ribosomes for Protein Synthesis. *Nature* **190**, 576–581 (1961).
214. Kapranov, P. *et al.* RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription. *Science* **316**, 1484–1488 (2007).
215. Mercer, T. R., Dinger, M. E. & Mattick, J. S. Long non-coding RNAs: insights into functions. *Nat Rev Genet* **10**, 155–159 (2009).
216. Bunch, H. Gene regulation of mammalian long non-coding RNA. *Mol Genet Genomics* **293**, 1–15 (2018).
217. Dykes, I. M. & Emanueli, C. Transcriptional and Post-transcriptional Gene Regulation by Long Non-coding RNA. *Genomics, Proteomics & Bioinformatics* **15**, 177–186 (2017).
218. Quinn, J. J. & Chang, H. Y. Unique features of long non-coding RNA biogenesis and function. *Nature Reviews Genetics* **17**, 47–62 (2016).
219. Sun, Q., Hao, Q. & Prasanth, K. V. Nuclear Long Noncoding RNAs: Key Regulators of Gene Expression. *Trends in Genetics* **34**, 142–157 (2018).

220. Brown, J., Hendrich, B. D. & Rupert, J. L. The Human XIST Gene: Analysis of a 17 kb Inactive X-Specific RNA That Contains Conserved Repeats and Is Highly Localized within the Nucleus. 16 (1992) doi:https://doi.org/10.1016/0092-8674(92)90520-M.
221. Brannan, C. I., Dees, E. C., Ingram, R. S. & Tilghman, S. M. The product of the H19 gene may function as an RNA. *Mol. Cell. Biol.* **10**, 28–36 (1990).
222. Rinn, J. L. *et al.* Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs. *Cell* **129**, 1311–1323 (2007).
223. Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigó, R. & Johnson, R. Towards a complete map of the human long non-coding RNA transcriptome. *Nature Reviews Genetics* (2018) doi:10.1038/s41576-018-0017-y.
224. Yang, L., Duff, M. O., Graveley, B. R., Carmichael, G. G. & Chen, L.-L. Genomewide characterization of non-polyadenylated RNAs. *Genome Biol* **12**, R16 (2011).
225. Hangauer, M. J., Vaughn, I. W. & McManus, M. T. Pervasive Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic Noncoding RNAs. *PLoS Genet* **9**, e1003569 (2013).
226. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Research* **22**, 1775–1789 (2012).
227. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
228. Wilusz, J. E. Long noncoding RNAs: Re-writing dogmas of RNA processing and stability. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1859**, 128–138 (2016).
229. Wucher, V. *et al.* FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Res* gkw1306 (2017) doi:10.1093/nar/gkw1306.
230. Ji, Z., Song, R., Regev, A. & Struhl, K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *eLife* **4**, e08890 (2015).
231. Brocchieri, L. Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Research* **33**, 3390–3400 (2005).
232. Geisler, S., Lojek, L., Khalil, A. M., Baker, K. E. & Coller, J. Decapping of Long Noncoding RNAs Regulates Inducible Genes. *Molecular Cell* **45**, 279–291 (2012).
233. 6 Non-coding RNA characterization. *Nature* nature28175 (2019) doi:10.1038/nature28175.
234. Wan, Y. *et al.* Systematic identification of intergenic long-noncoding RNAs in mouse retinas using full-length isoform sequencing. *BMC Genomics* **20**, 559 (2019).
235. Lagarde, J. *et al.* Extension of human lncRNA transcripts by RACE coupled with long-read high-throughput sequencing (RACE-Seq). *Nat Commun* **7**, 12339 (2016).
236. Rich, A., Kasha, M. & others. Horizons in biochemistry. *Eds. M. Kasha and B. Pullman, Academic Press, New York* 103 (1962).
237. Neveu, M., Kim, H.-J. & Benner, S. A. The "Strong" RNA World Hypothesis: Fifty Years Old. *Astrobiology* **13**, 391–403 (2013).
238. Gilbert, W. Origin of life: The RNA world. *Nature* **319**, 618–618 (1986).
239. Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N. & Altman, S. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* **35**, 849–857 (1983).

240. Westheimer, F. H. Biochemistry: Polyribonucleic acids as enzymes. *Nature* **319**, 534–536 (1986).
241. Breaker, R. R. Riboswitches and the RNA World. *Cold Spring Harbor Perspectives in Biology* **4**, a003566–a003566 (2012).
242. Cech, T. R. The RNA Worlds in Context. *Cold Spring Harbor Perspectives in Biology* **4**, a006742–a006742 (2012).
243. Higgs, P. G. & Lehman, N. The RNA World: molecular cooperation at the origins of life. *Nat Rev Genet* **16**, 7–17 (2015).
244. Schreiber, U. C. & Mayer, C. *The First Cell: The Mystery Surrounding the Beginning of Life*. (Springer International Publishing, 2020). doi:10.1007/978-3-030-45381-7.
245. Robertson, M. P. & Joyce, G. F. The Origins of the RNA World. *Cold Spring Harbor Perspectives in Biology* **4**, a003608–a003608 (2012).
246. Hezroni, H. *et al.* A subset of conserved mammalian long non-coding RNAs are fossils of ancestral protein-coding genes. *Genome Biol* **18**, 162 (2017).
247. Duret, L. The Xist RNA Gene Evolved in Eutherians by Pseudogenization of a Protein-Coding Gene. *Science* **312**, 1653–1655 (2006).
248. Elisaphenko, E. A. *et al.* A Dual Origin of the Xist Gene from a Protein-Coding Gene and a Set of Transposable Elements. *PLoS ONE* **3**, e2521 (2008).
249. Schmitz, J. & Brosius, J. Exonization of transposed elements: A challenge and opportunity for evolution. *Biochimie* **93**, 1928–1934 (2011).
250. Balakirev, E. S. & Ayala, F. J. Pseudogenes: Are They "Junk" or Functional DNA? *Annu. Rev. Genet.* **37**, 123–151 (2003).
251. Pink, R. C. *et al.* Pseudogenes: Pseudo-functional or key regulators in health and disease? *RNA* **17**, 792–798 (2011).
252. Mighell, A. J., Smith, N. R., Robinson, P. A. & Markham, A. F. Vertebrate pseudogenes. *FEBS Letters* **468**, 109–114 (2000).
253. Kutter, C. *et al.* Rapid Turnover of Long Noncoding RNAs and the Evolution of Gene Expression. *PLoS Genet* **8**, e1002841 (2012).
254. Ponjavic, J., Ponting, C. P. & Lunter, G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Research* **17**, 556–565 (2007).
255. Marques, A. C. & Ponting, C. P. Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biol* **10**, R124 (2009).
256. Zampetaki, A., Albrecht, A. & Steinhofel, K. Long Non-coding RNA Structure and Function: Is There a Link? *Front. Physiol.* **9**, 1201 (2018).
257. Hezroni, H. *et al.* Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species. *Cell Reports* **11**, 1110–1122 (2015).
258. Kirk, J. M. *et al.* Functional classification of long non-coding RNAs by k-mer content. *Nat Genet* **50**, 1474–1482 (2018).
259. Muret, K. *et al.* Long noncoding RNAs in lipid metabolism: literature review and conservation analysis across species. *BMC Genomics* **20**, 882 (2019).
260. Bush, S. J. *et al.* Cross-species inference of long non-coding RNAs greatly expands the ruminant transcriptome. *Genet Sel Evol* **50**, 20 (2018).
261. De Kumar, B. & Krumlauf, R. HOXs and lincRNAs: Two sides of the same coin. *Sci. Adv.* **2**, e1501402 (2016).

262. Ulitsky, I., Shkumatava, A., Jan, C. H., Sive, H. & Bartel, D. P. Conserved Function of lincRNAs in Vertebrate Embryonic Development despite Rapid Sequence Evolution. *Cell* **147**, 1537–1550 (2011).
263. The FANTOM Consortium. The Transcriptional Landscape of the Mammalian Genome. *Science* **309**, 1559–1563 (2005).
264. Noviello, T. M. R. *et al.* Detection of long non–coding RNA homology, a comparative study on alignment and alignment–free metrics. *BMC Bioinformatics* **19**, 407 (2018).
265. Amaral, P. P. *et al.* Genomic positional conservation identifies topological anchor point RNAs linked to developmental loci. *Genome Biol* **19**, 32 (2018).
266. Washietl, S., Kellis, M. & Garber, M. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Research* **24**, 616–628 (2014).
267. Necsulea, A. *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**, 635–640 (2014).
268. Kampa, D. Novel RNAs Identified From an In-Depth Analysis of the Transcriptome of Human Chromosomes 21 and 22. *Genome Research* **14**, 331–342 (2004).
269. Muret, K. *et al.* Long noncoding RNA repertoire in chicken liver and adipose tissue. *Genet Sel Evol* **49**, 6 (2017).
270. Le Béguec, C. *et al.* Characterisation and functional predictions of canine long non-coding RNAs. *Sci Rep* **8**, 13444 (2018).
271. Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development* **25**, 1915–1927 (2011).
272. Mele, M. *et al.* The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015).
273. de Goede, O. M. *et al.* Long non-coding RNA gene regulation and trait associations across human tissues. (2019) doi:10.1101/793091.
274. Scott, E. Y. *et al.* Identification of long non-coding RNA in the horse transcriptome. *BMC Genomics* **18**, 511 (2017).
275. Kryuchkova-Mostacci, N. & Robinson-Rechavi, M. A benchmark of gene expression tissue-specificity metrics. *Briefings in Bioinformatics* bbw008 (2016) doi:10.1093/bib/bbw008.
276. Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).
277. Ransohoff, J. D., Wei, Y. & Khavari, P. A. The functions and unique features of long intergenic non-coding RNA. *Nature Reviews Molecular Cell Biology* **19**, 143–157 (2017).
278. Kern, C. *et al.* Genome-wide identification of tissue-specific long non-coding RNA in three farm animal species. *BMC Genomics* **19**, 684 (2018).
279. Tang, Z. *et al.* Comprehensive analysis of long non-coding RNAs highlights their spatio-temporal expression patterns and evolutional conservation in Sus scrofa. *Sci Rep* **7**, 43166 (2017).
280. St. Laurent, G., Wahlestedt, C. & Kapranov, P. The Landscape of long noncoding RNA classification. *Trends in Genetics* **31**, 239–251 (2015).
281. Villegas, V. & Zaphiropoulos, P. Neighboring Gene Regulation by Antisense Long Non-Coding RNAs. *IJMS* **16**, 3251–3266 (2015).

282. Wight, M. & Werner, A. The functions of natural antisense transcripts. *Essays Biochem.* **54**, 91–101 (2013).
283. Luo, S. *et al.* Divergent lncRNAs Regulate Gene Expression and Lineage Differentiation in Pluripotent Cells. *Cell Stem Cell* **18**, 637–652 (2016).
284. Postepska-Igielska, A. *et al.* LncRNA Khps1 Regulates Expression of the Proto-oncogene SPHK1 via Triplex-Mediated Changes in Chromatin Structure. *Molecular Cell* **60**, 626–636 (2015).
285. Beltran, M. *et al.* A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition. *Genes & Development* **22**, 756–769 (2008).
286. Modarresi, F. *et al.* Inhibition of natural antisense transcripts in vivo results in gene-specific transcriptional upregulation. *Nat Biotechnol* **30**, 453–459 (2012).
287. Kotzin, J. J. *et al.* The long non-coding RNA Morrbid regulates Bim and short-lived myeloid cell lifespan. *Nature* **537**, 239–243 (2016).
288. Kambara, H. *et al.* Regulation of Interferon-Stimulated Gene BST2 by a lncRNA Transcribed from a Shared Bidirectional Promoter. *Front. Immunol.* **5**, 676 (2015).
289. Corney, B. P. A. *et al.* Regulatory Architecture of the Neuronal Cacng2/Tarpγ2 Gene Promoter: Multiple Repressive Domains, a Polymorphic Regulatory Short Tandem Repeat, and Bidirectional Organization with Co-regulated lncRNAs. *J Mol Neurosci* **67**, 282–294 (2019).
290. Lepoivre, C. *et al.* Divergent transcription is associated with promoters of transcriptional regulators. *BMC Genomics* **14**, 914 (2013).
291. Volders, P.-J. *et al.* LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Research* **41**, D246–D251 (2013).
292. Ma, L. *et al.* LncBook: a curated knowledgebase of human long non-coding RNAs. *Nucleic Acids Research* **47**, D128–D134 (2019).
293. Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet* **47**, 199–208 (2015).
294. Li, A. *et al.* ALDB: A Domestic-Animal Long Noncoding RNA Database. *PLoS ONE* **10**, e0124003 (2015).
295. Fang, S. *et al.* NONCODEV5: a comprehensive annotation database for long non-coding RNAs. *Nucleic Acids Research* **46**, D308–D314 (2018).
296. Koufariotis, L. T., Chen, Y.-P. P., Chamberlain, A., Vander Jagt, C. & Hayes, B. J. A catalogue of novel bovine long noncoding RNA across 18 tissues. *PLoS ONE* **10**, e0141225 (2015).
297. Clark, E. L. *et al.* A high resolution atlas of gene expression in the domestic sheep (Ovis aries). 38 (2017).
298. Loda, A. & Heard, E. Xist RNA in action: Past, present, and future. *PLoS Genet* **15**, e1008333 (2019).
299. Hacisuleyman, E. *et al.* Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. *Nat Struct Mol Biol* **21**, 198–206 (2014).
300. Jain, A. K. *et al.* LncPRESS1 Is a p53-Regulated LncRNA that Safeguards Pluripotency by Disrupting SIRT6-Mediated De-acetylation of Histone H3K56. *Molecular Cell* **64**, 967–981 (2016).

301. Peterlin, B. M., Brogie, J. E. & Price, D. H. 7SK snRNA: a noncoding RNA that plays a major role in regulating eukaryotic transcription: 7SK: a regulatory snRNA. *WIREs RNA* **3**, 92–103 (2012).
302. Latos, P. A. *et al.* An in vitro ES cell imprinting model shows that imprinted expression of the Igf2r gene arises from an allele-specific expression bias. *Development* **136**, 437–448 (2009).
303. Engreitz, J. M. *et al.* Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* **539**, 452–455 (2016).
304. Romero-Barrios, N., Legascue, M. F., Benhamed, M., Ariel, F. & Crespi, M. Splicing regulation by long noncoding RNAs. *Nucleic Acids Research* **46**, 2169–2184 (2018).
305. Gong, C. & Maquat, L. E. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3′ UTRs via Alu elements. *Nature* **470**, 284–288 (2011).
306. Lee, S. *et al.* Noncoding RNA NORAD Regulates Genomic Stability by Sequestering PUMILIO Proteins. *Cell* **164**, 69–80 (2016).
307. Jégu, T., Aeby, E. & Lee, J. T. The X chromosome in space. *Nat Rev Genet* **18**, 377–389 (2017).
308. Clemson, C. M., McNeil, J. A., Willard, H. F. & Lawrence, J. B. XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure. *The Journal of Cell Biology* **132**, 259–275 (1996).
309. Creamer, K. M. & Lawrence, J. B. *XIST* RNA: a window into the broader role of RNA in nuclear chromosome architecture. *Phil. Trans. R. Soc. B* **372**, 20160360 (2017).
310. Engreitz, J., Lander, E. S. & Guttman, M. RNA Antisense Purification (RAP) for Mapping RNA Interactions with Chromatin. in *Nuclear Bodies and Noncoding RNAs* (eds. Nakagawa, S. & Hirose, T.) vol. 1262 183–197 (Springer New York, 2015).
311. Mariner, P. D. *et al.* Human Alu RNA Is a Modular Transacting Repressor of mRNA Transcription during Heat Shock. *Molecular Cell* **29**, 499–509 (2008).
312. Arcot, S. S., Wang, Z., Weber, J. L., Deininger, P. L. & Batzer, M. A. Alu Repeats: A Source for the Genesis of Primate Microsatellites. *Genomics* **29**, 136–144 (1995).
313. Espinoza, C. A., Goodrich, J. A. & Kugel, J. F. Characterization of the structure, function, and mechanism of B2 RNA, an ncRNA repressor of RNA polymerase II transcription. *RNA* **13**, 583–596 (2007).
314. Allen, T. A., Von Kaenel, S., Goodrich, J. A. & Kugel, J. F. The SINE-encoded mouse B2 RNA represses mRNA transcription in response to heat shock. *Nat Struct Mol Biol* **11**, 816–821 (2004).
315. C. Quaresma, A. J., Bugai, A. & Barboric, M. Cracking the control of RNA polymerase II elongation by 7SK snRNP and P-TEFb. *Nucleic Acids Res* **44**, 7527–7539 (2016).
316. Ebisuya, M., Yamamoto, T., Nakajima, M. & Nishida, E. Ripples from neighbouring transcription. *Nat Cell Biol* **10**, 1106–1113 (2008).
317. Birger, Y., Shemer, R., Perk, J. & Razin, A. The imprinting box of the mouse Igf2r gene. *Nature* **397**, 84–88 (1999).
318. Latos, P. A. *et al.* Airn Transcriptional Overlap, But Not Its lncRNA Products, Induces Imprinted Igf2r Silencing. *Science* **338**, 1469–1472 (2012).
319. Spector, D. L. & Lamond, A. I. Nuclear Speckles. *Cold Spring Harbor Perspectives in Biology* **3**, a000646–a000646 (2011).

320. Mao, Y. S., Zhang, B. & Spector, D. L. Biogenesis and function of nuclear bodies. *Trends in Genetics* **27**, 295–306 (2011).
321. Tripathi, V. *et al.* The Nuclear-Retained Noncoding RNA MALAT1 Regulates Alternative Splicing by Modulating SR Splicing Factor Phosphorylation. *Molecular Cell* **39**, 925–938 (2010).
322. Zhang, B. *et al.* Identification and Characterization of a Class of MALAT1 -like Genomic Loci. *Cell Reports* **19**, 1723–1738 (2017).
323. Kalluri, R. & Weinberg, R. A. The basics of epithelial-mesenchymal transition. *J. Clin. Invest.* **119**, 1420–1428 (2009).
324. Hu, J. *et al.* Long Noncoding RNA EGFR-AS1 Promotes Cell Proliferation by Increasing EGFR mRNA Stability in Gastric Cancer. *Cell Physiol Biochem* **49**, 322–334 (2018).
325. Salmena, L., Poliseno, L., Tay, Y., Kats, L. & Pandolfi, P. P. A ceRNA Hypothesis: The Rosetta Stone of a Hidden RNA Language? *Cell* **146**, 353–358 (2011).
326. Mockly, S. & Seitz, H. Inconsistencies and Limitations of Current MicroRNA Target Identification Methods. in *MicroRNA Target Identification* (ed. Laganà, A.) vol. 1970 291–314 (Springer New York, 2019).
327. Denzler, R., Agarwal, V., Stefano, J., Bartel, D. P. & Stoffel, M. Assessing the ceRNA Hypothesis with Quantitative Measurements of miRNA and Target Abundance. *Molecular Cell* **54**, 766–776 (2014).
328. Kleaveland, B., Shi, C. Y., Stefano, J. & Bartel, D. P. A Network of Noncoding Regulatory RNAs Acts in the Mammalian Brain. *Cell* **174**, 350-362.e17 (2018).
329. Baccarini, A. *et al.* Kinetic Analysis Reveals the Fate of a MicroRNA following Target Regulation in Mammalian Cells. *Current Biology* **21**, 369–376 (2011).
330. Ameres, S. L. *et al.* Target RNA-Directed Trimming and Tailing of Small Silencing RNAs. *Science* **328**, 1534–1539 (2010).
331. Tichon, A. *et al.* A conserved abundant cytoplasmic long noncoding RNA modulates repression by Pumilio proteins in human cells. *Nat Commun* **7**, 12209 (2016).
332. Miller, M. A. & Olivas, W. M. Roles of Puf proteins in mRNA degradation and translation: Puf proteins in mRNA degradation and translation. *WIREs RNA* **2**, 471–492 (2011).
333. Bordas, A., Tixier-Boichard, M. & Merat, P. Direct and correlated responses to divergent selection for residual food intake in Rhode island red laying hens. *British Poultry Science* **33**, 741–754 (1992).
334. Tixier-Boichard, M., Boichard, D., Groeneveld, E. & Bordas, A. Restricted Maximum Likelihood Estimates of Genetic Parameters of Adult Male and Female Rhode Island Red Chickens Divergently Selected for Residual Feed Consumption. *Poultry Science* **74**, 1245–1252 (1995).
335. Gabarrou, J.-F., Géraert, P.-A., Picard, M. & Bordas, A. Diet-Induced Thermogenesis in Cockerels Is Modulated by Genetic Selection for High or Low Residual Feed Intake. *The Journal of Nutrition* **127**, 2371–2376 (1997).
336. El-Kazzi, M., Bordas, A., Gandemer, G. & Minvielle, F. Divergent selection for residual food intake in Rhode Island Red egg-laying lines: Gross carcase composition, carcase adiposity and lipid contents of tissues. *British Poultry Science* **36**, 719–728 (1995).

337. Tixier, M., Bordas, A. & Merat, P. Divergent selection for residual feed intake in laying hens: effects on growth and fatness. in *Leanness in Domestic Birds Genetic, Metabolic and Hormonal Aspects* 420 (B. Leclercq & C.C. Whitehead, 1988).
338. Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigó, R. & Johnson, R. Towards a complete map of the human long non-coding RNA transcriptome. *Nat Rev Genet* **19**, 535–548 (2018).
339. Burge, S. *et al.* Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation. *Database* **2012**, bar068–bar068 (2012).
340. Pedruzzi, I. *et al.* HAMAP in 2015: updates to the protein family classification and annotation system. *Nucleic Acids Research* **43**, D1064–D1070 (2015).
341. Lizio, M. *et al.* Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol* **16**, 22 (2015).
342. Lizio, M. *et al.* Systematic analysis of transcription start sites in avian development. *PLOS Biology* **15**, e2002887 (2017).
343. Jiménez-Badillo, S. E., Oviedo, N., Hernández-Guzmán, C., González-Mariscal, L. & Hernández-Sánchez, J. Catsper1 promoter is bidirectional and regulates the expression of a novel lncRNA. *Sci Rep* **7**, 13351 (2017).
344. Jehl, F. *et al.* An integrative atlas of chicken long non-coding genes and their annotations across 25 tissues. *Scientific Reports* **10**, 20457 (2020).
345. Genome Aggregation Database Consortium *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
346. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).
347. Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).
348. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297–1303 (2010).
349. Poplin, R. *et al. Scaling accurate genetic variant discovery to tens of thousands of samples*. http://biorxiv.org/lookup/doi/10.1101/201178 (2017) doi:10.1101/201178.
350. Panousis, N. I., Gutierrez-Arcelus, M., Dermitzakis, E. T. & Lappalainen, T. Allelic mapping bias in RNA-sequencing is not a major confounder in eQTL studies. 8 (2014).
351. 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 (Methodological)* **57**, 289–300 (1995).
352. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2019).
353. Stier, A. *et al.* Avian erythrocytes have functional mitochondria, opening novel perspectives for birds as animal models in the study of ageing. *Front Zool* **10**, 33 (2013).
354. Désert, C. *et al.* Transcriptomes of whole blood and PBMC in chickens. *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics* **20**, 1–9 (2016).
355. Popadin, K., Gutierrez-Arcelus, M., Dermitzakis, E. T. & Antonarakis, S. E. Genetic and Epigenetic Regulation of Human lincRNA Gene Expression. *The American Journal of Human Genetics* **93**, 1015–1026 (2013).

356. Primeau, M. & Lamarche-Vane, N. Coup d'œil sur les petites GTPases Rho. *Med Sci (Paris)* **24**, 157–162 (2008).
357. Kim, K. R. *et al.* Structural and Biophysical Analyses of Human N-Myc Downstream-Regulated Gene 3 (NDRG3) Protein. *Biomolecules* **10**, 90 (2020).
358. Grobas, S., Mendez, J., De Blas, C. & Mateos, G. Laying hen productivity as affected by energy, supplemental fat, and linoleic acid concentration of the diet. *Poultry Science* **78**, 1542–1551 (1999).
359. Harms, R. H., Russell, G. B. & Sloan, D. R. Performance of Four Strains of Commercial Layers With Major Changes in Dietary Energy. *The Journal of Applied Poultry Research* **9**, 535–541 (2000).
360. Murugesan, G. R. & Persia, M. E. Validation of the effects of small differences in dietary metabolizable energy and feed restriction in first-cycle laying hens. *Poultry Science* **92**, 1238–1243 (2013).
361. Di Marzo, V. & Matias, I. Endocannabinoid control of food intake and energy balance. *Nature Neuroscience* **8**, 585–589 (2005).
362. Bermudez-Silva, F. J., Viveros, M. P., McPartland, J. M. & Rodriguez de Fonseca, F. The endocannabinoid system, eating behavior and energy homeostasis: The end or a new beginning? *Pharmacology Biochemistry and Behavior* **95**, 375–382 (2010).
363. Jehl, F. *et al.* Chicken adaptive response to low energy diet: main role of the hypothalamic lipid metabolism revealed by a phenotypic and multi-tissue transcriptomic approach. *BMC Genomics* **20**, 1033 (2019).
364. Smith, C. L. & Eppig, J. T. The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *WIREs Syst Biol Med* **1**, 390–399 (2009).
365. Sun, Y. *et al.* The identification of 14 new genes for meat quality traits in chicken using a genome-wide association study. *BMC Genomics* **14**, 458 (2013).
366. Huyghe, J. R. Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nature Genetics* **45**, 7 (2013).
367. Kolic, J. *et al.* Differential Effects of Voclosporin and Tacrolimus on Insulin Secretion From Human Islets. *Endocrinology* **161**, bqaa162 (2020).
368. Prentice, K. J., Saksi, J. & Hotamisligil, G. S. Adipokine FABP4 integrates energy stores and counterregulatory metabolic responses. *J. Lipid Res.* **60**, 734–740 (2019).
369. Chung, J. Y., Ain, Q. U., Song, Y., Yong, S.-B. & Kim, Y.-H. Targeted delivery of CRISPR interference system against Fabp4 to white adipocytes ameliorates obesity, inflammation, hepatic steatosis, and insulin resistance. *Genome Res.* **29**, 1442–1452 (2019).
370. Thompson, B. R., Mazurkiewicz-Muñoz, A. M., Suttles, J., Carter-Su, C. & Bernlohr, D. A. Interaction of Adipocyte Fatty Acid-binding Protein (AFABP) and JAK2: AFABP/aP2 AS A REGULATOR OF JAK2 SIGNALING. *J. Biol. Chem.* **284**, 13473–13480 (2009).
371. Villeneuve, J. *et al.* Unconventional secretion of FABP4 by endosomes and secretory lysosomes. *Journal of Cell Biology* **217**, 649–665 (2018).
372. Zeng, J., Sauter, E. R. & Li, B. FABP4: A New Player in Obesity-Associated Breast Cancer. *Trends in Molecular Medicine* **26**, 437–440 (2020).

373. Nakamura, R. *et al.* Serum fatty acid-binding protein 4 (FABP4) concentration is associated with insulin resistance in peripheral tissues, A clinical study. *PLoS ONE* **12**, e0179737 (2017).
374. Wang, Q., Guan, T., Li, H. & Bernlohr, D. A. A novel polymorphism in the chicken adipocyte fatty acid-binding protein gene (FABP4) that alters ligand-binding and correlates with fatness. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* **154**, 298–302 (2009).
375. Kim, A. *et al.* Synonymous variants in holoprosencephaly alter codon usage and impact the Sonic Hedgehog protein. *Brain* **143**, 2027–2038 (2020).
376. Brule, C. E. & Grayhack, E. J. Synonymous Codons: Choose Wisely for Expression. *Trends in Genetics* **33**, 283–297 (2017).
377. Gilbert, H. *et al.* Sélection pour la consommation alimentaire moyenne journalière résiduelle chez le porc : impacts sur les caractères et défis pour la filière. *INRA Prod. Anim.* **30**, 439–454 (2018).
378. Herd, R. M. *et al.* Genetic divergence in residual feed intake affects growth, feed efficiency, carcass and meat quality characteristics of Angus steers in a large commercial feedlot. *Anim. Prod. Sci.* **58**, 164 (2018).
379. Kelly, A. K. *et al.* Effect of divergence in residual feed intake on feeding behavior, blood metabolic variables, and body composition traits in growing beef heifers1. *Journal of Animal Science* **88**, 109–123 (2010).
380. Zhuo, Z., Lamont, S. J., Lee, W. R. & Abasht, B. RNA-Seq Analysis of Abdominal Fat Reveals Differences between Modern Commercial Broiler Chickens with High and Low Feed Efficiencies. *PLoS ONE* **10**, e0135810 (2015).
381. Roux, P.-F. *et al.* Combined QTL and selective sweep mappings with coding SNP annotation and cis-eQTL analysis revealed PARK2 and JAG2 as new candidate genes for adiposity regulation. *G3: Genes, Genomes, Genetics* **5**, 517–529 (2015).
382. Sheng, Z., Pettersson, M. E., Honaker, C. F., Siegel, P. B. & Carlborg, Ö. Standing genetic variation as a major contributor to adaptation in the Virginia chicken lines selection experiment. *Genome Biol* **16**, 219 (2015).
383. Hager, R., Cheverud, J. M. & Wolf, J. B. Relative contribution of additive, dominance and imprinting effects to phenotypic variation in body size and growth between divergent selection lines of mice. *Evolution* **63**, 1118–1128 (2009).
384. Park, Y.-M., Myers, M. & Vieira-Potter, V. J. Adipose tissue inflammation and metabolic dysfunction: role of exercise. *Mo Med* **111**, 65–72 (2014).
385. Greenberg, A. S. & Obin, M. S. Obesity and the role of adipose tissue in inflammation and metabolism. *The American Journal of Clinical Nutrition* **83**, 461S-465S (2006).
386. Wellen, K. E. & Hotamisligil, G. S. Obesity-induced inflammatory changes in adipose tissue. *J. Clin. Invest.* **112**, 1785–1788 (2003).
387. Kusminski, C. M. & Scherer, P. E. Mitochondrial dysfunction in white adipose tissue. *Trends in Endocrinology & Metabolism* **23**, 435–443 (2012).
388. Sun, K., Kusminski, C. M. & Scherer, P. E. Adipose tissue remodeling and obesity. *J. Clin. Invest.* **121**, 2094–2101 (2011).
389. Choe, S. S., Huh, J. Y., Hwang, I. J., Kim, J. I. & Kim, J. B. Adipose Tissue Remodeling: Its Role in Energy Metabolism and Metabolic Disorders. *Front. Endocrinol.* **7**, (2016).

390. Maarouf, M. *et al.* Identification of lncRNA-155 encoded by MIR155HG as a novel regulator of innate immunity against influenza A virus infection. *Cellular Microbiology* **21**, (2019).
391. Ng, P. C. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research* **31**, 3812–3814 (2003).
392. Sim, N.-L. *et al.* SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Research* **40**, W452–W457 (2012).
393. Exome Aggregation Consortium *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

DOCTORAT ECOLOGIE
BRETAGNE GEOSCIENCES
LOIRE AGRONOMIE ALIMENTATION

l'institut Agro
agriculture • alimentation • environnement

AGRO CAMPUS OUEST

Titre : Étude de la composante génétique de l'efficience alimentaire (EA) chez des lignées de poules pondeuses divergentes pour l'EA en utilisant la technologie RNA-seq

Mots clés : efficience alimentaire, poule pondeuse, *cis*-régulation, ARN long non-codant, RNA-seq, trace de sélection

Résumé : L'efficience alimentaire (EA) est un important caractère d'intérêt agronomique. La variation de l'EA dans une population est principalement due à la variation d'expression de gènes, elle-même due à des variants qui les régulent en *cis*. Cette thèse avait trois objectifs interconnectés.
Le 1er était d'étudier les tissus et gènes impliqués dans la différence d'EA entre deux lignées de pondeuses divergentes pour ce caractère. Nous avons étudié pour cela les transcriptomes de quatre tissus (tissu adipeux, sang, hypothalamus et foie) à l'aide des possibilités offertes par le RNA-seq, technologie de séquençage des ARN.
Le 2e objectif était de contribuer à l'annotation fonctionnelle du génome de la poule, en enrichissant son annotation en gènes d'ARN longs non-codants, d'importants régulateurs de l'expression. Nous avons également détecté les SNP par RNA-seq, étape nécessaire à la mise en place d'un *pipeline* permettant l'étude de l'expression allèle-spécifique, un marqueur des *cis*-régulations.
Le 3e objectif combinait les résultats des travaux précédents afin d'identifier des gènes candidats causaux de la variation d'EA, en raison d'un variant impactant la structure de l'ARN ou protéine associée, ou bien d'un variant *cis*-régulateur.

Title : Study of the genetic component of feed efficiency (FE) in layer lines divergent for FE using the RNA-seq technology

Keywords : feed efficiency, laying hen, *cis*-regulation, long non-coding RNA, RNA-seq, selective sweep

Abstract: Feed efficiency (FE) is an important trait of agronomical interest. The variation of FE in a population is mainly due to the variation of gene expression, itself due to variants that regulate them in *cis*. This thesis had three interconnected objectives.
The 1st was to study the tissues and genes involved in the difference in FE between two different strains of layers divergent for this trait. To this end, we studied transcriptomes from four tissues (adipose tissue, blood, hypothalamus and liver) using the possibilities offered by RNA-seq, an RNA sequencing technology.
The 2nd objective was to contribute to the functional annotation of the chicken genome by enriching its annotation with long non-coding RNA genes, important expression regulators. We also detected SNPs by RNA-seq, a necessary step in the establishment of a pipeline for the study of allele-specific expression, a marker of cis-regulation.
The 3rd objective combined the results of previous work to identify candidate genes that cause FE variation due to a variant impacting the structure of the RNA or associated protein, or a cis-regulatory variant.