THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Institut National Polytechnique de Toulouse (Toulouse INP)

Discipline ou spécialité :

Pathologie, Toxicologie, Génétique et Nutrition

Présentée et soutenue par :

Mme ÉMILIE DELPUECH

le mardi 6 avril 2021

Titre :

Architecture génétique de l'efficacité alimentaire chez le porc en croissance : exploration de données génétique et transcriptomique issues de lignées divergentes sur la consommation moyenne journalière

---

Ecole doctorale :

Sciences Ecologiques, Vétérinaires, Agronomiques et Bioingénieries (SEVAB)

Unité de recherche :

Génétique, Physiologie et Systèmes d'Elevage (GENPHYSE)

Directeur(s) de Thèse :

MME JULIETTE RIQUET

Rapporteurs :

MME ELISABETH DUVAL, INRA TOURS

MME PASCALE LE-ROY, INRA LE RHEU

MME SANDRINE LAGARRIGUE, AGROCAMPUS OUEST

Membre(s) du jury :

MME CHRISTELE ROBERT-GRANIE, INRA TOULOUSE, Président

M. DOMINIQUE HAZARD, INRA TOULOUSE, Membre

MME JULIETTE RIQUET, INRA TOULOUSE, Membre

Table des matières

Table des matières ____ ***2***

Lexique ____ ***8***

Table des figures ____ ***10***

Table des tableaux ____ ***18***

Introduction générale ____ ***22***

Chapitre 1 : Synthèse bibliographique ____ ***24***

I. L'efficacité alimentaire, un caractère d'intérêt pour les animaux d'élevage ____ 24

1. Un contexte et un marché qui évoluent ____ 24
   a. La consommation mondiale de viande de porc ne cesse d'augmenter ____ 24
   b. Une demande des consommateurs pour un élevage plus respectueux ____ 26
   c. Concurrence entre alimentation animale et humaine ____ 27
2. Les mesures de la consommation et du comportement alimentaire ____ 28
   a. Améliorer l'efficacité alimentaire des animaux ____ 29
   b. Le caractère CMJR pour l'efficacité alimentaire ____ 30
3. Création de lignées divergentes pour la RFI ____ 34
   a. Les dispositifs expérimentaux de sélection pour la CMJR ____ 34
      i. Dispositif de sélection divergente INRAE ____ 34
      ii. Dispositif de sélection divergente de l'Iowa State University (ISU, USA) ____ 36
   b. Impact de la sélection divergente sur les autres caractères des animaux sélectionnés ____ 38
      i. Sur les caractères d'efficacité alimentaire et de composition corporelle ____ 38
      ii. Sur la qualité de viande ____ 40
      iii. Sur le comportement alimentaire ____ 40
      iv. Pour la digestibilité ____ 41
      v. Sur les processus métaboliques des porcs ____ 41
4. Vers la prédiction génomique de l'efficacité alimentaire ____ 42

II. Cartographie des régions QTL pour l'efficacité alimentaire ____ 45

1. Structure de la variabilité du génome de porc et cartographie ____ 45
   a. Evolution des outils d'analyse de la variabilité du génome ____ 45
   b. Importance du déséquilibre de liaison pour la cartographie ____ 46
   c. Les études d'associations, outil clé d'identification de régions QTL ____ 49
2. L'imputation de génotypes pour affiner la localisation de variants ____ 52

a. Principe 52
b. Les différents domaines d'utilisation de l'imputation 55
3. Des régions génomiques identifiées pour l'efficacité alimentaire 57
a. Cartographie de QTL pour la RFI chez le porc 57
b. Cartographie chez les autres espèces de rente 58
III. Apport des données fonctionnelles à la cartographie de QTL 59
1. Les données transcriptomiques et leurs utilisations 59
a. Evolution des technologies d'analyse du transcriptome 59
b. Les analyses transcriptome appliquées à la RFI chez le porc 61
i. Caractérisation physiologique 61
ii. Utilisation comme phénotype (eQTL) 64
2. L'annotation fonctionnelle pour la cartographie de régions génomiques 65
a. De l'évolution des connaissances vers leurs mises en commun 65
b. Le consortium FAANG pour les animaux de rentes 68
c. L'ontologie génétique et ses outils 70
d. Approches novatrices pour la caractérisation des régions génomiques en ségrégation à l'aide de données fonctionnelles 73
Chapitre 2 : Identification de régions QTL pour la CMJR à partir des lignées divergentes INRAE 78
I. Introduction 78
II. Stratégie 79
III. Identification et correction des erreurs de pedigree 81
1. Contrôle des erreurs de parentés 81
2. Les erreurs de sexe 83
3. Les erreurs de générations et de lignées 84
IV. Article : Identification des régions génomiques affectant les caractères de production des porcs sélectionnés de manière divergente pour l'efficacité alimentaire 86
Abstract 87
Background 87
Methods 88
Ethic statement 88
Results 93
Discussion 102
Conclusion 106
Additionnal file 1 106

Additionnal file 2 ____ 107
Additionnal file 3 ____ 107
Declarations ____ 108
References ____ 109

V. Comparaison des régions QTL identifiées pour la CMJR avec celles publiées pour d'autres espèces de rentes ____ 112

VI. Discussion ____ 114

Chapitre 3 : Intégration de données génétiques, transcriptomiques et d'annotations fonctionnelles ____ 117

I. Introduction ____ 118

II. Analyse d'enrichissement à partir de données transcriptomiques obtenues sur les lignées CMJR, Gondret et al. 2017 ____ 119

1. Le dispositif et les données disponibles ____ 119
2. Analyse d'enrichissement ____ 121
   a. Analyses réalisées par Gondret et al ____ 121
   b. Analyses réalisées à l'aide de l'outil GSEA ____ 123

III. Combinaison des données génétiques et transcriptomiques ____ 125

1. Les gènes différentiellement exprimés ____ 125
   a. Les gènes positionnés sur le génome du porc ____ 125
   b. Sélection des DEG les plus significatifs ____ 126
2. Recherche de candidats positionnels dans les régions QTL parmi les DEG sélectionnés ____ 126

IV. Etudes d'enrichissement à partir des régions QTL ____ 129

1. Les études d'enrichissement avec GSEA ____ 129
   a. Application de GSEA sur des résultats de GWAS ____ 129
      i. Les options de GSEA choisies ____ 129
      ii. Adaptation de la librairie de termes GO et de la liste à analyser ____ 130
   b. Résultats obtenus à partir d'une liste de gènes ____ 132
2. Une approche par fenêtre avec l'outil GSEA ____ 135
   a. Les études d'enrichissement sur des fenêtres génomiques ____ 136
   b. Identification des voies métaboliques caractéristiques du dispositif ____ 137

V. Analyse simultanée des 3 types de données ____ 141

1. Détection de clusters de DEG ____ 141
   a. Proportion de DEG par fenêtre ____ 142
   b. Moyenne d'expression par fenêtre ____ 143

2. Des régions génétiques "sous sélection" 144
a. Estimation de l'évolution des fréquences alléliques 145
b. Estimation de l'évolution de la structure du DL 147
3. Etude d'enrichissement 151
a. Liste des fenêtres sélectionnées pour l'analyse 151
b. Analyses GSEA : intégration des données génétiques et transcriptomiques 152

VI. Discussion 156

1. Vers une meilleure compréhension fonctionnelle des lignées divergentes ? 156
2. Limites de la combinaison des données génétiques et transcriptomiques 158

Conclusion et perspectives 163

Références 171

Annexe 1 189

Annexe 2 193

Annexe 3 197

Annexe 4 212

Lexique

- **ACP** : Analyse en Composantes Principales

- **ADN** : Acide Désoxyribonucléique

- **BD** : Basse Densité

- **CMJ** (en anglais, Feed Intake FI) : Consommation Moyenne Journalière

- **CMJR** (Residual Feed Intake, RFI) : Consommation Moyenne Journalière Résiduelle

- **DAC** : Distributeur Automatique de Concentrés

- **DEG** : « *Differentially Expressed Gene* », soit les gènes différentiellement exprimés

- **DL** (Linkage Disequilibrium, LD) : Déséquilibre Liaison

- **EA** (Feed Efficiency, FE): Efficacité Alimentaire

- **EBV** : « Estimated breeding values », soit les valeurs d'élevage estimées

- **ELD** (backfat, BF): Epaisseur de Lard Dorsal

- **FAO** : « *Food and Agriculture Organization of the United Nations* », soit l'Organisation des Nations Unies pour l'alimentation et l'agriculture

- **FDR** : « *False Discovery Rate* », soit le taux d'erreur

- **GEBV** : « Genomic Estimated Breeding Value »

- **GO** : « *Gene Ontology* », soit l'ontologie génétique

- **GSEA** : « *Gene Set Enrichment Analysis* », soit l'analyse de l'enrichissement des ensembles de gènes

- **GMQ** : Gain de poids Moyen Quotidien

- **GWAS** : « *Genome-Wide Association Studies* », soit l'étude d'association pangénomique

- **HD** (High Density) : Haute Densité

- **IC** : indice de consommation

- **IFIP** : Institut de la filière porcine

- **INRAE** : Institut national de recherche pour l'agriculture, l'alimentation et l'environnement

- **ISU** : Iowa State University (USA)

- **IQV** (Meat Quality Index, MQI) : Indice de Qualité de Viande

- **MAF** : Fréquence Alélique Mineure

- **MD** (Medium Density) : Moyenne Densité

- **NGS** : « Next Generation Sequencing » pour les séquenceurs de nouvelle génération

- **PCR** : « Polymerase Chain Reaction »

- **PV** : Poids Vif

- **PMM** : Poids Métabolique Moyen

- **PRAT** : « Perirenal Adipose Tissue », soit le tissu adipeux périrénal

- **QTL** : « *Quantitative Trait Loci* », soit les loci des caractères quantitatifs

- **QTN** : « *Quantitative Trait Nucleotide* », soit les nucléotides des caractères quantitatifs

- **RDT** : Rendement de carcasse

- **SCAT** : « Subcutaneous Adipose Tissue », soit le tissu adipeux sous cutané

- **SNP** : « *Single Nucleotide Polymorphism* », soit le polymorphisme d'un seul nucléotide

- **SSC** : « *Sus Scrofa Chromosome »*

- **ssGBLUP** : « *Single-step Genomic Best Linear Unbiased Prediction* »

- **TMP** : Taux de Muscle des Pièces à l'abattoir

- **TVM** : Teneur en Viande Maigre

- **WGS** : « W*hole Genome Shotgun sequences* », soit le séquençage *shotgun* (littéralement séquençage "fusil de chasse") permettant de séquencer des brins d'ADN aléatoires.

Table des figures

Figure 1 : Représentation du poids de viandes de volaille, porcine, ovine et bovine (en kg par personne et par an) sur les 6 continents et dans le monde pour les années 2016 à 2018 et également une estimation de cette consommation de viande en 2028. Note: la consommation par habitant est exprimée en poids au détail (OCDE-FAO, 2019).....24

Figure 2 : Distribution de la production mondiale de viande, estimée en millions de tonnes de 1961 à 2018, selon les différents types d'élevages (avec principalement les volailles, les porcs, les bovins et les ovins). Note : la production total de viande inclut les abattages commerciaux et personnels. Les données sont récupérées en termes de poids carcasses sans les abats et les graisses d'abattage. Poultry = poulet ; Pigmeat = viande porcine ; Beef and Buffalo = bovins ; Sheep and Goat = ovins ; Goose and guinea fowl = oie et pintade ; Camel = camélidé ; Horse = cheval ; Duck =canard ; Wild game = gibier sauvage. (Ritchie and Roser, 2017).....25

Figure 3 : Consommation totale de viande (en tonne par an) de 1970 à 2019 en France avec une distinction entre principales fillières d'élevages français, le poulet (Chicken, jaune), le porc (Pork, orange), le boeuf (Beef, violet) et le mouton (Mutton, bleu) (FranceAgriMer, 2020). 25

Figure 4 : Les enjeux environnementaux des élevages porcins divisés en deux parties selon la priorité des élevages (IFIP, 2011) 26

Figure 5 : Composition moyenne de l'alimentation des animaux d'élevages en France, avec pour unité le kg de masse sèche, pour les 4 principales filières françaises (Volailles, Porcins, Ovins, Bovins) (GIS, 2017) 27

Figure 6 : Coût de revient en euro par kg de carcasse de la filière porcine dans divers pays dont la France (rouge). Les coûts de revient dans chaque pays ont été répartis en 4 grandes catégories : l'alimentation (orange), les autres charges opérationnelles (jaune), la main d'œuvre (vert) et les amortissements et frais financiers (bleu) (IFIP, 2020) 28

Figure 7 : Exemple de la composition de la carcasse du porc standard (en poids) en partant du poids à vif d'un porc, avec les différents parties du porcs valorisées ou perdues au cours de l'abattage et de la découpe des divers produits issus de la carcasse des porcs (GIS, 2017).....28

Figure 8 : L'énergie ingérée par les porcs se répercute dans 5 grandes voies : l'énergie emmagasinée, l'énergie utilisée pour leur entretien et l'activité par les individus, l'énergie nécessaire au maintien de la chaleur des individus, l'énergie pour les fèces et l'énergie de l'urine et du méthane (Phocas et al., 2014). 29

Figure 9 : Schéma de la consommation résiduelle d'aliments comme différence entre la consommation d'aliments observée et la consommation estimée, basée sur le GMQ et l'ELD du porc. L'écart à la droite représente la CMJR : les animaux qui se trouvent au-dessus de la droite ont une CMJR>0 (les moins efficaces, CMJR+), alors que les animaux positionnés en dessous de la droite caractérisent les animaux à CMJR<0 (les plus efficaces, CMJR-) (Azarpajouh et al., 2017).....31

Figure 10 : Schéma de la sélection divergente sur le caractère CMJR des porcs Large White (Lignées INRAE) avec la création de 2 parités (les candidats à la sélection et les animaux réponses) pour chacune des lignées divergentes (CMJR+ et CMJR-). Cette sélection divergente a été réalisée sur 10 générations, de la G0 à la G9. La P1 se caractérise par l'accouplement des 6 mâles sélectionnés à la génération précédente (parmi 96 mâles testés) avec environ 40 femelles (non testées et remplacées par une fille à la génération suivante). La P2 se compose d'environ 48 mâles et 48 femelles issus de l'accouplement des mêmes parents.....35

Figure 11 : Schéma du dispositif de sélection des porcs Yorkshire, (a) avec la création de la lignée CMJR- (Lignées ISU) et la création de deux parités (P1 en vert ; P2 en orange) caractérisées pour la P1 par les mâles testés et sélectionnés en même temps que les femelles non testées par rapport à leurs valeurs génétiques pour la CMJR, et pour la P2 par l'accumulation de données phénotypiques pour la CMJR de sœurs ou demi-sœurs des reproducteurs sélectionnés. (b) La lignée contrôle a été obtenue par sélection aléatoire des reproducteurs jusqu'à la génération G4 (vert clair) puis une lignée CMJR+ a été sélectionnée sur base de leurs valeurs génétiques pour la CMJR (vert). 37

Figure 12 : Réponses à la sélection sur la Consommation Moyenne Journalière Résiduelle (CMJR) pour les caractères de croissance, de consommation et de composition de carcasse de la génération G0 à la génération G9, exprimées en écart-types génétiques des caractères. CMJR+ = lignée à CMJR élevée, moins efficace ; CMJR– = lignée à CMJR faible, plus efficace ; CMJ = consommation moyenne journalière ; GMQ = Gain Moyen Quotidien ; IC = Indice de Consommation ; ELD = Épaisseur de Lard Dorsal ; IQV = Indice de Qualité de la Viande ; TVM = Taux de Viande Maigre de la carcasse estimé par combinaison linéaire des poids de morceaux (Gilbert et al., 2017a).....39

Figure 13 : Réponses directes et corrélées à la sélection sur la consommation alimentaire résiduelle (CMJR) basées sur la valeur génétique moyenne et la comparaison directe des lignées en génération 4 (4-Phen). Les valeurs génétiques moyennes et les barres d'erreurs standards sont basées uniquement sur les verrats CMJR-. 4-Phen : Différences phénotypiques des lignées (contrôle et CMJR-) basées sur la comparaison phénotypique des lignées de la génération 4 des lignées ISU ; ADFI = consommation moyenne journalière ; FE (efficacité alimentaire) = ADG/ADFI ; BF = épaisseur de lard dorsal ; LMA = surface de muscle de la longe ; IMF = teneur en gras intramusculaire (Cai et al., 2008). 39

Figure 14 : Le déséquilibre de liaison (DL) moyen pour toutes les populations à différentes distances présentées en kb. La relation entre le DL prédit (LD_{ij}) et la distance (paires de bases) est indiquée par race et par région génomique : cible 1 (a), cible 2 (b) et cible 3 (c). Les races chinoises sont représentées par des lignes pointillées, les races européennes par des lignes pleines, et le sanglier par une ligne pleine épaisse (Amaral et al., 2008). 48

Figure 15 : Distribution du déséquilibre de liaison via le r^2 de 0 à 10 Mb dans 4 populations de porcs Américaines (Badke et al., 2014) 49

Figure 16 : Nombre de publications par an utilisant les méthodes GWAS pour tous types d'organismes vivants (a) et sur les GWAS pour les animaux d'élevages (b) (source : pubmed Décembre 2020) 51

Figure 17 : Schéma illustrant les analyses réalisées des études GWAS jusqu'à l'identification de QTN sur le génome ciblé.....52

Figure 18 : Schéma explicatif du processus d'imputation, avec (a) des individus possédant un panel de génotypes complets et (b) des individus avec des données génotypiques manquantes. (c) La

combinaison de la population de référence et de la population candidate permet de retrouver les haplotypes correspondants. Les haplotypes identifiés dans la population candidate à l'imputation, ont été colorés en fonction des haplotypes de référence auxquels ils correspondent. Au final, (d) la population candidate obtient un génotype complet après imputation. (Marchini and Howie, 2010) ..53

Figure 19 : Représentation de la qualité d'imputation selon différentes tailles de population, N=60 individus (rouge), N=200 individus (orange) ou N=500 individus (bleu). Les courbes illustrent l'impact de la proportion de marqueurs dont les génotypes sont imputés avec précision (r^2 élevé entre les génotypes imputés et réels). (Li et al., 2009) 54

Figure 20 : (a) Taux de génotype correct par rapport aux fréquences et (b) corrélation entre les valeurs des vrais génotypes et des génotypes imputés par rapport aux fréquences alléliques des SNPs pour l'imputation moyenne densité (54 000 marqueurs) à haute densité (777 000 marqueurs) en utilisant différentes méthodes d'imputation : IMPUTE2, Beagle, findhap, AlphaImpute et FImpute. AlphaBea signifie que les génotypes manquants après imputation par AlphaImpute ont été imputés par Beagle.(Ma et al., 2013) 55

Figure 21 : Schéma explicatif de l'imputation dans le cadre de la prédiction génomique où les descendants sont génotypés sur une puce BD, et les parents et grand-parents sur une puce moyenne ou haute densité. PGF = Grand-père paternel ; PGM = Grand-mère paternelle ; MGF = Grand-père maternel ; MGM = Grand-mère maternelle ; Father = Père ; Mother = Mère. (Burdick et al., 2006)....57

Figure 22 : Positionnement des QTL identifiés pour 4 caractères caractéristiques de l'efficacité alimentaire (CMJR, CMJ, IC et GMQ) et uniquement identifiés dans des études sur l'analyse de la CMJR chez des porcs en croissance. Seuls les 18 autosomes de porcs ont été représentés. 58

Figure 23 : Résultat général de l'étude d'annotation fonctionnelle de la réponse du foie et du tissu adipeux au jeûne, en identifiant les voies biologiques KEGG surreprésentées (P<0,05) en gènes différentiellement exprimés par Lkhagvadorj (Lkhagvadorj et al., 2009) 63

Figure 24 : Le réseau clé de gènes et d'ARNmi s'est avéré être exprimé différemment dans le muscle squelettique des porcs CMJR- par rapport à celui des porcs CMJR+. Le réseau a été réalisé à l'aide de Cytoscape. (Jing et al., 2015).....64

Figure 25 : Les éléments génomiques fonctionnels identifiés par la phase I d'ENCODE. Les méthodes indiquées sont utilisées pour identifier différents types d'éléments fonctionnels dans le génome humain. (Consortium, 2004).....66

Figure 26 : Acquisition de données omics à différentes échelles pour obtenir des données sur l'ADN, l'ARN, les protéines et les métabolites, afin de mieux comprendre la biologie de caractères complexes. (Sun and Hu, 2016).....66

Figure 27 : Accumulation de nouveaux types d'analyses au cours des trois phases d'ENCODE. (Snyder et al., 2020) 67

Figure 28 : L'Institut national de recherche sur le génome humain (NHGRI) a identifié une liste de publications qui ont utilisé les données ENCODE. Les publications communautaires (« Community publications ») sont identifiées par des recherches automatisées pour la citation des numéros

d'adhésion ENCODE, des documents phares ENCODE ou des ressources telles que HaploReg et RegulomeDB et manuellement pour déterminer si les données ENCODE ont été effectivement utilisées. Les documents du consortium (« ENCODE consortium ») sont identifiés par des recherches automatisées dans PubMed pour les publications qui ont été soutenues au moins en partie par des prix ENCODE. (Moore et al., 2020) 68

Figure 29 : Annotation fonctionnelle des génomes par l'analyse d'ensembles de données obtenues par différentes approches de séquençage a permis de construire des cartes complètes d'annotation du génome, comprenant des éléments qui agissent au niveau des protéines et de l'ARN et des éléments de régulation. (Giuffra and Tuggle, 2019) 69

Figure 30 : Représentation d'un ensemble de termes GO décrite via un graphe, où chaque terme GO est un nœud, et les relations entre les termes sont des arêtes entre les nœuds. Par exemple, le terme de processus biologique "processus biosynthétique de l'hexose" a deux parents, "processus métabolique de l'hexose" et "processus biosynthétique des monosaccharides" (http://geneontology.org/docs/ontology-documentation/). 71

Figure 31 : Illustration de la méthode GSEA. (a) Un ensemble de données d'expression triées par rapport au différentiel d'expression obtenu entre les deux phénotypes étudiés et la carte thermique correspondante à ces valeurs, c'est-à-dire l'emplacement des gènes d'un ensemble S dans la liste triée. (b) Graphique de la somme courante pour S dans l'ensemble de données, y compris l'emplacement du score d'enrichissement maximal (ES). (Subramanian et al., 2005) 72

Figure 32 : La génétique des systèmes combine la génomique génétique, l'association génétique et les analyses génomiques pour construire l'inférence causale de génotype à phénotype. L'approche intègre les informations sur le génotype (g), de la ~omique (o) et les données sur le phénotype (p) et peut être utilisée pour construire un réseau et déduire la causalité des différents variants. (van der Sijde et al., 2014) 74

Figure 33 : Principe d'analyse des données génétiques issus du dispositif de sélection divergente pour la CMJR. (a) Trois puces génétiques ont été utilisées : une puce MD 60K (MD-V1, Illumina Porcine SNP60v2 BeadChip), une puce MD 70K (MD-V2, Illumina Porcine HD Array GGP) et une puce HD (Affymetrix Axiom Porcine HD Genotyping Array). Pour ces 3 jeux de données, un contrôle qualité sur les données génétiques a été appliqué et un contrôle du pedigree a également été réalisé grâce aux données génétiques MD. (b) Avec l'ensemble des données deux étapes d'imputation ont été réalisées, une première imputation pour homogénéiser les données MD et une seconde imputation pour avoir l'ensemble des individus avec des génotypes HD. Enfin un génotype parental moyen a été affecté aux animaux de la seconde parité grâce aux données génétiques imputés en HD pour les animaux P1. (c) Au final, des GWAS ont été réalisées avec les génotypes parentaux moyens et des phénotypes directement mesurées sur ces individus. 80

Figure 34 : Différents génotypes de descendants lors du croisement d'un mâle (génotypes 0/1/2 encadré bleu) avec une femelle (génotypes 0/1/2 encadré rouge) sont possibles. Néanmoins, si les deux parents sont codés 1, tous les génotypes sont possibles (cases rouges). 81

Figure 35 : Résultat du test de parenté pour les 1 589 individus dont les deux parents ont été génotypés. Les individus bleus ne possèdent que quelques erreurs mendéliennes pour les 42 800 SNPs testés, en revanche les individus rouges sont détectés à problème car plus de 1 000 erreurs mendéliennes ont été détectées. 82

Figure 36 : Schéma des chromosomes X et Y qui possèdent une portion chromosomique homologue entre les deux chromosomes sexuels X et Y (PAR en vert) et une portion spécifique à chacun des chromosomes (spécifique X/Y en beige). 83

Figure 37 : Représentation des valeurs d'hétérozygotie des variants du chromosome X avec l'identification de la PAR et de la portion spécifique pour les mâles génotypés (a) sur la puce 60K et (b) sur la puce 70K. 84

Figure 38 : Représentation de l'hétérozygotie des individus génotypés avec la distinction des femelles (rouge) et des mâles (bleu) au vu des valeurs de fréquences alléliques 84

Figure 39 : Caryotype représentant les 18 autosomes et où chaque chromosome est caractérisé par le nombre de fenêtres de 1Mb. La position des gènes présents dans les QTL identifiés en bovin (rond violet) et en poulet (rond rose) ont été reportés sur le côté droit des chromosomes et les régions QTL détectées pour la CMJR dans mon analyse sont représentées sur le côté gauche des chromosomes (rond orange) 113

Figure 40 : Diagramme de Venn de la combinaison des 3 types de données (données génétiques : « QTL », données transcriptomiques : « DEG » et les données d'annotation : « Annotation ») pris en compte dans notre étude, afin d'affiner l'identification de gènes candidats pour la cartographie fine des régions QTL en lien avec l'efficacité alimentaire. 118

Figure 41 : Barplot du nombre de DEG par tissu qui sont issus de l'analyse des données transcriptomiques présenté dans l'article de Gondret et al. (Gondret et al., 2017) dont le muscle (« Muscle »), le foie (« Liver »), le tissu adipeux périrénal (« PRAT ») et le tissu adipeux sous cutané (« SCAT »). 120

Figure 42 : Distribution du nombre de DEG par rapport à leur différentiel d'expression 120

Figure 43 : Diagramme de Venn des DEG identifiés dans les 4 tissus étudiés, le muscle (« Muscle »), le foie (« Liver ») et les 2 tissus adipeux (« PRAT » et « SCAT ») d'après les résultats d'analyse des données transcriptomiques présentés dans l'article de Gondret et al. (Gondret et al., 2017). 121

Figure 44 : Exemple d'apparentement entre les termes GO identifiés par Gondret et al. (en bleu) et les termes GO identifiés avec GSEA (en vert). Trois seuils de significativité sont indiqués, FDR<0,01 (vert foncé), 0,01 <FDR<0,05 (vert intermédiaire) et 0,05<FDR<0,25 (vert clair). 124

Figure 45 : Représentation sur le caryotype du génome de porc de la localisation de l'ensemble des gènes exprimés de la puce (pourpre) et des DEG (bleu) identifiés par Gondret et al. Les DEG avec un facteur de différentiel d'expression d'au moins 2 (dans l'une ou l'autre lignée) sont représentés sur le caryotype par des croix encerclées, à gauche des DEG. 126

Figure 46 : Représentation schématique de le l'analyse d'enrichissement réalisé. Trois termes GO (1, 2 et 3) sont définis par une liste de gènes (A, B, C, D, E, F) positionnés sur différents chromosomes ou dans une même région génomique. Ces gènes sont alors ordonnés en fonction de la valeur du $-\log_{10}$(p-value) attribué à la fenêtre où le gène est localisé. 131

Figure 47 : Comparaison des deux valeurs d'ordonnancement : (a) distribution des maximum de -log10(p-value) (en rouge) et des somme des -log10(p-value) (en bleu). (a) Dot Plot pour l'ensemble des gènes de la valeur du maximum des -log10(p-value) versus la somme des -log10(p-value).....132

Figure 48 : Schéma de l'analyse GSEA effectuée sur la liste des gènes du génome de porc avec deux listes de gènes ordonnés soit par le maximum des –log10(p-value) de la fenêtre, soit par la somme des –log10(p-value) de la fenêtre. Deux listes de termes GO résultats sont obtenus.....133

Figure 49 : Localisation le long du génome des fenêtres contenant les différents gènes du terme GO "cornification". Pour chaque fenêtre le nombre de gènes dans le GO est indiqué sur l'axe des Y. Le gradient de couleur indique la valeur du $-\log_{10}$(p-value) maximum associé aux gènes de la fenêtre. 135

Figure 50 : Principe de transformation des termes GO caractérisant un ensemble de gènes possédant des caractéristiques fonctionnelles communes (a), en termes GO caractérisant des régions génomiques où sont positionnés des gènes avec des caractéristiques communes (b).....136

Figure 51 : Représentation schématique de le l'analyse d'enrichissement réalisé à partir des fenêtres. Le schéma est identique à celui présenté en figure 46 mais les gènes des termes Go et la liste soumise à l'analyse GSEA sont cette fois transformés en fenêtres ; les gènes localisés dans une même fenêtre (gènes A et B) n'apparaissent qu'une seule fois dans les termes GO et dans la liste ordonnée. 137

Figure 52 : Schéma de l'analyse GSEA effectuée sur la liste des fenêtres du génome de porc avec deux listes de gènes ordonnés soit par le maximum des –log10(p-value) de la fenêtre, soit par la somme des –log10(p-value) de la fenêtre. Deux listes de termes GO résultats sont obtenus.....138

Figure 53 : Apparentement entre les termes GO identifiés à partir des fenêtres QTL (encadrés jaune) ou à partir de la liste de l'ensemble des DEG identifiés par Gondret & al. Dans les deux cas, les analyses ont été réalisées à l'aide de GSEA. 139

Figure 54 : Distribution du nombre de DEG par fenêtre de 1Mb pour les 2 271 fenêtres du génome de porc. 142

Figure 55 : Distribution de la proportion de DEG (en pourcentage, %) dans les fenêtres de 1 Mb contenant au moins 3 DEG. 143

Figure 56 : Plot de la moyenne du différentiel par rapport au nombre de gènes dans la fenêtre pour les intervalles génomiques contenant au moins 12, 6 et 3 DEG dans les proportions de DEG]25-50] (violet),]50-75] (vert) et]75-100] (jaune) respectivement. En complément, pour chaque proportion, la moyenne globale des moyennes du différentiel de chaque fenêtre a été reportée via les lignes en pointillés de couleurs similaires à chaque proportion. Pour finir, la courbe représentant au mieux l'ensemble des points a été représentée en gris. 144

Figure 57 : Schéma de détection de régions en sélection par l'analyse des fréquences alléliques et du DL de manière commune. L'évolution des fréquences alléliques après plusieurs générations de sélection sont représentées dans la partie haute du schéma (a) et l'évolution du DL dans une région génomique en sélection est visualisée via des cartes triangulaires de chaleur.....145

Figure 58 : Représentation de 4 profils différents d'évolution de fréquences alléliques entre les générations G0 et en G7 pour les deux lignées divergentes. Chaque paire de graphes représente une fenêtre d'1 Mb, et chaque point un SNP. Le graphe supérieur correspond aux fréquences d'un allèle de référence en G0, le graphe inférieur, la différence de fréquence entre les générations G0 et G7 pour chaque lignée (CMJR+ en violet, CMJR- en vert). La valeur moyenne de la différence des fréquences alléliques entre les lignées en G7 a été reportée sur chaque représentation en rouge. Les figures représentent (a) deux exemples de régions dont les fréquences ont fortement divergé entre les lignées et (b) deux exemples d'absence d'évolution entre lignées. 147

Figure 59 : Représentation des blocs de DL via des cartes triangulaires de chaleur du r^2, en G0 et G7 pour les 2 lignées (CMJR- et CMJR+) et pour un intervalle génomique donné. Le DL se visualise par un dégradé de rouge, directement lié aux valeurs de r^2 entre chaque marqueur de la région génomique (rouge : $r^2 = 1$; blanc : $r^2 = 0$).....148

Figure 60 : Schéma des 4 profils d'évolution de la structure génétique des fenêtres sous l'effet de la sélection en fonction de la structure du DL et des fréquences alléliques aux SNP de la fenêtre. 149

Figure 61 : Représentation de la proportion de composantes principales perdues entre la G0 et la G7, en fonction de la moyenne des différences alléliques en G7 entre les deux lignées pour les 103 fenêtres contenant les DEG sélectionnés (DEG avec différentiel > 2 et/ou des clusters de DEG). Les fenêtres violettes correspondent à celles possédant un cluster de DEG, les bleues sont celles contenant un DEG à différentiel fort et les vertes correspondent aux fenêtres contenant un DEG à différentiel fort et un cluster de DEG. Les pointillés rouges représentent les seuils choisis pour distingués les 4 profils d'évolution.....150

Figure 62 : Apparentement entre 4 termes GO identifiés pour les caractères du temps d'imbibition du muscle et du pH 24h adducteur.....154

Figure 63 : Apparentement entre les termes GO identifiés dans les analyses successives : en bleu les termes GO identifiés par Gondret et al. (Gondret et al., 2017) à l'aide de l'outil DAVID, en orange les termes GO identifiés à partir de la même liste de DEG et l'outil GSEA (partie I), en jaune les termes GO identifiés à partir des régions QTL (partie II) et en vert les termes GO identifiés en combinant les informations génomiques et transcriptomiques (partie IV). 155

Table des tableaux

Table 1 : Estimations de l'héritabilité pour la CMJR issues de la littérature (Hoque and Suzuki, 2009) 33

Table 2 : Héritabilités(h^2) et corrélations génétiques avec la CMJR (r_g) des caractères mesurés pour les quatre populations (Saintilan et al., 2012) 34

*Table 3 : Paramètres génétiques estimés (h^2=héritabilité ; ρ_g = corrélation génétique avec la CMJR ; σ_g = écart-type génétique ; σ_p = écart-type phénotypique), et réponses génétiques à la sélection en génération G9 dans les lignées CMJR+ et CMJR-, ainsi que le niveau de signification de la différence (p-value) : p-value pour la différence entre les moyennes des moindres carrés des valeurs génétiques pour les lignées CMJR- et CMJR+ de la génération G9 avec *** = P<0,001, * = 0,01<P<0,05, t = 0,05<P<0,10. (Gilbert et al., 2017b) 36*

Table 4 : Paramètres génétiques estimés pour le dispositif des lignées ISU. RFI = CMJR ; ADFI = CMJ ; ADG = GMQ ; FE = efficacité alimentaire (EA) ; BF = ELD ; LMA = superficie de la longe ; IMF = teneur en gras intramusculaire. (Cai et al., 2008).....38

Table 5 : Table récapitulative des précisions de prédictions des caractères liés à l'efficacité alimentaire calculées sur des populations de porcs. La corrélation entre les prédictions génomiques et le phénotype a été divisée par la racine carrée de l'héritabilité et certaines valeurs (b) ont été arrondies à partir de la fiabilité rapportée dans les publications. (Zhang et al., 2018).....43

Table 6 : Etat de l'art des différentes puces de génotypage porcines présentent sur le marché. 46

Table 7 : Liste des principaux processus biologiques mis en évidence pour chaque tissu et présentés dans l'article de Gondret et al. (Gondret et al., 2017).....122

Table 8 : Présentation du nombre de gènes présents sur le support de la puce transcriptomique ainsi que le nombre de ces gènes positionnés sur le génome de porc V11.2. Au sein de ces deux groupes de gènes, deux catégories de gènes ont été prises en compte pour notre analyse : les gènes exprimés de la puce et parmi ces gènes ceux qui sont différentiellement exprimés entre les lignées (CMJR- et CMJR+). 125

Table 9 : Liste des DEG présentant un différentiel d'expression > 2 et localisés dans ou à proximité d'une région QTL. 127

Table 10 : Table résultat de GSEA correspondant au 41 termes GO les plus significatifs avec le caractère pour lequel chaque GO présente la plus grande significativité. Seulement 3 valeurs statistiques issues de l'analyse GSEA sont représentées dans cette table : ES (Enrichment Score), NES (Normalize Enrichment Score) et FDR.q.val (False Discovery Rate).....134

Table 11 : Liste des termes GO identifiés (FDR<0,25) avec les analyses GSEA effectuées sur les 2 271 fenêtres. 138

Table 12 : Liste des fenêtres positionnées dans des régions QTL ou à quelques méga bases de celles-ci. 151

Table 13 : Liste des fenêtres avec évolution de structure génétique entre les lignées 152

Table 14 : Liste des termes GO identifiés (FDR<0,25) avec les analyses GSEA effectuées sur les 2 271 fenêtres sélectionnées à partir des données génétiques et transcriptomiques.....153

Introduction générale

L'efficacité alimentaire est un caractère majeur pour l'économie de la filière porcine compte tenu des coûts des aliments qui représentent à eux seuls plus de 2/3 des coûts de production totaux. Depuis de nombreuses années, la sélection est considérée comme un levier d'action pour améliorer l'efficacité des animaux et réduire les pertes. Parmi les mesures disponibles pour l'évaluation de l'efficacité alimentaire, la consommation moyenne journalière résiduelle (CMJR) est un caractère qui reflète la différence entre la consommation observée et la consommation théorique estimée afin de répondre aux besoins de production et d'entretien. A INRAE depuis plus de 20 ans, une sélection divergente pour une meilleure (CMJR-) et une moindre efficacité alimentaire (CMJR+) a été effectuée à partir de porcs Large-White (LW) pendant dix générations. En utilisant ces lignées sélectionnées, l'objectif principal de cette thèse est d'approfondir nos connaissances sur les bases génétiques et biologiques de la CMJR chez les porcs en croissance. Dans un premier chapitre je me suis attachée à faire une synthèse bibliographique pour rappeler le contexte dans lequel se situe mon travail de thèse et faire un état de l'art sur les approches de cartographie de QTL et l'apport des connaissances fonctionnelles à la compréhension de l'architecture génétique des caractères.

Le second chapitre est destiné à présenter les différentes analyses génétiques menées sur les lignées divergentes INRAE pour identifier des régions génomiques affectant la CMJR, et des caractères de production corrélés à la CMJR via des analyses d'association. Les données de génotypage ont été acquises à l'aide de puces de SNP de moyenne densité (MD) pour l'ensemble des individus reproducteurs des deux lignées divergentes et 32 individus fondateurs ont été génotypés à l'aide d'une puce haute densité (HD). Grace à des analyses d'imputation nous avons pu reconstruire les génotypes de 570 447 marqueurs de la puce HD pour l'ensemble des reproducteurs. Ces génotypes ont alors été utilisés pour attribuer un génotype moyen à l'ensemble des portées de descendants phénotypés pour les caractères d'intérêt (dont le CMJR). Ainsi des analyses GWAS ont été réalisées à partir d'un dispositif de 2 426 porcs phénotypés pour 24 caractères et disposant après prédiction des génotypes reconstruits à l'aide des informations généalogiques. Ces études ont été réalisées dans chaque lignée indépendamment ou en combinant les deux lignées. L'ensemble des travaux réalisés dans ce chapitre ont été soumis pour publication dans la revue « Genetics Selection Evolution » (GSE) ; l'article est actuellement en révision.

Dans un troisième chapitre, je présenterai le travail destiné à affiner les régions identifiées à l'aide de données fonctionnelles complémentaires. Cette partie de ma thèse a été initiée lors d'une mobilité internationale réalisée dans le cadre du parcours doctoral de l'École Internationale de Recherche

d'Agreenium (EIR-A)[1]. J'ai réalisé un séjour à l'Iowa State University (ISU) dans l'Iowa, du 1er Novembre 2019 au 15 Mars 2020 dans l'équipe du Dr Christopher Tuggle. Au cours de cette mobilité j'ai eu la possibilité de prendre en main l'outil GSEA que j'ai utilisé pour la réalisation d'études d'enrichissement en combinant les résultats du premier chapitre de cette thèse et différents types de données fonctionnelles. J'ai ainsi pu exploiter un jeu de données transcriptomiques disponible sur 4 tissus pour un lot d'individus issu du même dispositif. Des études d'enrichissement successives ont été réalisées à l'aide de l'outils GSEA, (*i*) en confrontant uniquement la liste des gènes différentiellement exprimée (DEG) à la base de données « Molecular Signatures Database » (MSigDB), (*ii*) en recherchant dans cette même base de données les voies métaboliques enrichies en gènes présents dans les régions QTL et (*iii*) en cherchant à combiner les données génétiques et transcriptomiques.

Enfin, dans une dernière partie, je proposerai des pistes de travaux complémentaires à réaliser pour poursuivre ce travail de thèse.

[1] Agreenium est l'alliance de la formation et la recherche pour l'agriculture, l'alimentation, l'environnement et la santé globale, dans le but de rassembler, sur la base du volontariat, la majeure partie des établissements publics d'enseignement supérieur et des organismes de recherche placés sous la tutelle du ministre chargé de l'agriculture, dont la cellule de coordination est hébergée par INRAE. Ainsi une sélection d'une quarantaine de doctorants sur la base des sujets de thèse et de la motivation de chacun est réalisée chaque année, afin d'effectuer deux séminaires sur des sujets d'actualités et un séjour à l'international d'une durée minimum de trois mois durant le doctorat.

Chapitre 1 : Synthèse bibliographique

I. L'efficacité alimentaire, un caractère d'intérêt pour les animaux d'élevage

1. Un contexte et un marché qui évoluent

a. La consommation mondiale de viande de porc ne cesse d'augmenter

L'homme trouve principalement les protéines nécessaires à son développement et au bon fonctionnement de son organisme dans les aliments carnés. D'après la FAO la consommation de viande dans le monde, toutes espèces confondues, a progressé de 2,3% par an depuis les années 60 pour atteindre 325 Mt en 2019. La demande en viande devrait continuer à progresser (Figure 1) et atteindre 470 Mt en 2050. De façon schématique, la consommation de viande est répartie en 3 tiers entre la viande de porc, de volaille et de ruminant (bovins et ovins). Il est à noter qu'en dehors de la viande, la production mondiale d'œufs (volaille) et de lait (bovins, ovins et caprins) représente respectivement 66 Mt et 852 Mt en 2019 (OCDE and Organisation des Nations Unies pour l'alimentation et l'agriculture, 2020).

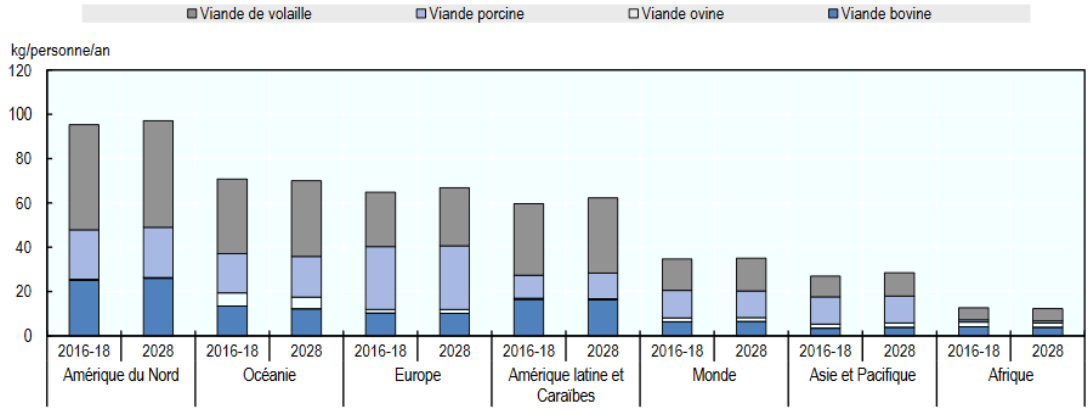

Figure 1 : Représentation du poids de viandes de volaille, porcine, ovine et bovine (en kg par personne et par an) sur les 6 continents et dans le monde pour les années 2016 à 2018 et également une estimation de cette consommation de viande en 2028. Note: la consommation par habitant est exprimée en poids au détail (OCDE-FAO, 2019).

Cette projection mondiale cache cependant des différences selon les régions du monde. L'augmentation de viande devrait s'accélérer dans les pays en développement en raison d'une forte croissance démographique (Figure 2), ainsi que de la hausse du niveau de vie qui s'accompagne d'un essor des classes moyennes (Amérique du Sud et en Asie). En revanche, dans les pays développés la consommation tend à diminuer sans que cette diminution n'équilibre l'augmentation de la demande dans d'autres pays (Ritchie and Roser, 2017). La consommation de viande porcine présente également des différences en fonction des régions du monde. Contrairement aux autres viandes, elle est quasiment nulle dans certaines régions, sa consommation étant soumise à des interdits dans celles-ci (OCDE-FAO, 2019), mais ses coûts de production relativement faibles et une forte affinité pour cette

viande en Asie devraient néanmoins entrainer une augmentation de sa consommation dans les décennies à venir.

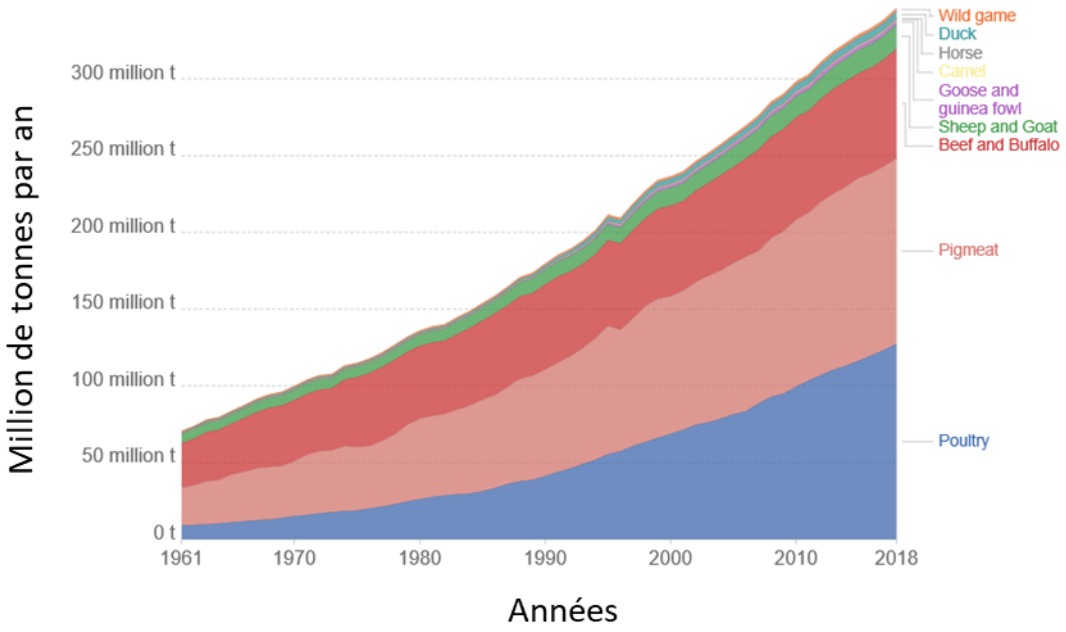

Figure 2 : Distribution de la production mondiale de viande, estimée en millions de tonnes de 1961 à 2018, selon les différents types d'élevages (avec principalement les volailles, les porcs, les bovins et les ovins). Note : la production total de viande inclut les abattages commerciaux et personnels. Les données sont récupérées en termes de poids carcasses sans les abats et les graisses d'abattage. Poultry = poulet ; Pigmeat = viande porcine ; Beef and Buffalo = bovins ; Sheep and Goat = ovins ; Goose and guinea fowl = oie et pintade ; Camel = camélidé ; Horse = cheval ; Duck =canard ; Wild game = gibier sauvage. (Ritchie and Roser, 2017)

En France la viande de porc reste la viande la plus consommée, malgré un recul constaté depuis presque 20 ans (baisse de 7 % de la consommation individuelle de viande de porc entre 1999 et 2019) (Figure 3). La consommation française de viande s'élève désormais à 5,7 millions de tonnes équivalent carcasses (tec) en 2019, soit un repli de 0,9 % par rapport à 2018 (Agreste, 2020).

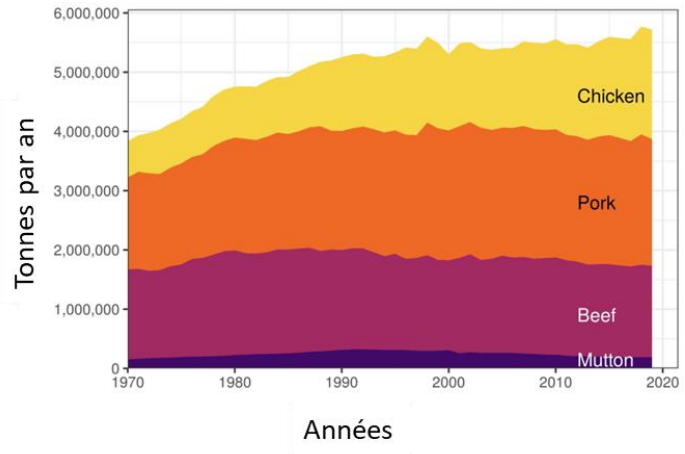

Figure 3 : Consommation totale de viande (en tonne par an) de 1970 à 2019 en France avec une distinction entre principales fillières d'élevages français, le poulet (Chicken, jaune), le porc (Pork, orange), le boeuf (Beef, violet) et le mouton (Mutton, bleu) (FranceAgriMer, 2020).

Le porc bénéficie d'un prix peu élevé par rapport à la viande de bœuf et de mouton (OCDE-FAO, 2019), ainsi que d'une importante diversité de produits, ce qui peut expliquer son maintien en tant que première viande consommée en France malgré une baisse de sa consommation (FranceAgriMer, 2020). Néanmoins, comme pour d'autres produits carnés, la diminution de la consommation de viande de porc en France ou en Europe est liée à de nouveaux facteurs sociétaux dont il est difficile d'estimer l'impact à plus long terme sur la consommation.

b. Une demande des consommateurs pour un élevage plus respectueux

Dans une grande partie des pays développés, le niveau de consommation de viande est arrivé à un plateau et de nouveaux comportements alimentaires, bien que minoritaires, sont néanmoins observés (végétarisme, véganisme), liés à une attention portée par les consommateurs aux conditions d'élevage des animaux, aux modes de production et à leurs impacts sur l'environnement.

L'impact environnemental majeur de la production porcine est associé à la concentration géographique des élevages généralement observée dans de nombreux pays. En France les ¾ de la production est localisée dans l'ouest de la France (Bretagne, Normandie, Pays de la Loire), la Bretagne représentant à elle seule 58,4% de la production en 2019 (+1,2% par rapport à 2009). Cette intensification toujours plus forte de l'élevage de porcs se heurte à une contestation voire une opposition de plus en plus marquée de la part de la société dont la sensibilité à la qualité de l'environnement est désormais un critère important (Delanoue and Roguet, 2015; Roguet et al., 2017). Pour les élevages de porcs, les enjeux environnementaux prioritaires sont la quantité de nitrate et de phosphore présente dans les excréments et l'impact des émissions d'ammoniac et de gaz à effet de serre (Figure 4) (IFIP, 2011).

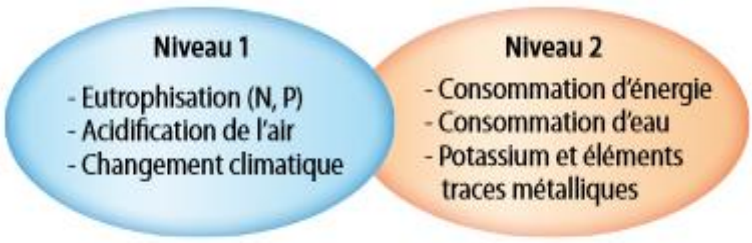

Figure 4 : Les enjeux environnementaux des élevages porcins divisés en deux parties selon la priorité des élevages (IFIP, 2011)

Le nitrate et le phosphore sont retrouvés dans les excréments des porcs et lors de l'épandage du lisier, ces molécules contaminent les sols et les cours d'eau. Des études sont menées pour imaginer la structure des élevages de demain et des solutions sont recherchées pour réduire l'impact environnemental tout en étant durables pour la gestion des contraintes réglementaires, le revenu des éleveurs, l'organisation du travail et le bien-être des animaux. Différents leviers d'actions sont en cours

d'évaluation pour revoir les modalités de gestion des effluents d'élevage, la conduite des exploitations, la réduction de la consommation des énergies et de l'eau, la gestion sanitaire des élevages ou les stratégies alimentaires des animaux (Wall et al., 2010; Ali et al., 2018; Soleimani and Gilbert, 2020).

c. *Concurrence entre alimentation animale et humaine*

La croissance de la population mondiale associée à une augmentation forte de la demande en denrées alimentaires entrainent progressivement une concurrence accrue pour l'utilisation des ressources telles que les terres arables pour la production de l'alimentation humaine ou animale, ou l'utilisation de l'eau.

Les produits agricoles cultivés et utilisés dans le monde servent traditionnellement à nourrir les Hommes et les animaux ainsi qu'à certains usages non-alimentaires. Dans un contexte d'augmentation de la demande alimentaire mondiale, les productions animales font débat car le rendement de la transformation des végétaux par les animaux est généralement faible, et l'alimentation des animaux d'élevages comporte une part de céréales consommables par l'Homme (Figure 5). On estime qu'en moyenne 2,5 à 10 kg de protéines végétales sont nécessaires pour produire 1 kg de protéines animales (GIS, 2017).

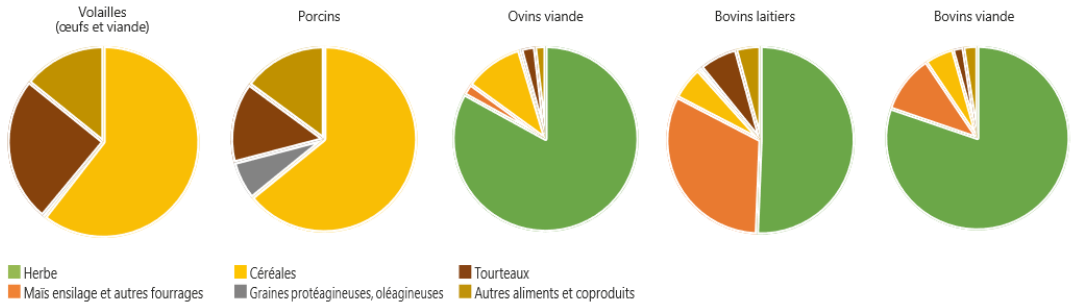

Figure 5 : Composition moyenne de l'alimentation des animaux d'élevages en France, avec pour unité le kg de masse sèche, pour les 4 principales filières françaises (Volailles, Porcins, Ovins, Bovins) (GIS, 2017)

En France, les céréales représentent respectivement environ 65% et 60% de l'alimentation des porcs et des volailles (chair et ponte) contrairement à l'élevage des ruminants dont les rations sont composées d'au moins 50% de fourrage (Dronne, 2018). Cette forte dépendance de l'alimentation des monogastriques aux céréales induit que la part de l'alimentation représente à elle seule près de 70% des coûts de production dans les élevages (Figure 6). Une des voies d'amélioration porte sur l'utilisation de coproduits de céréales qui ne sont pas consommables par l'homme et l'utilisation de nouvelles sources protéiques autres que le soja (colza, pulpes de betterave, luzerne…). L'objectif affiché est de réduire la concurrence avec l'alimentation humaine et de gagner en autonomie nationale (a minima européenne) afin de réduire l'impact de la fluctuation des cours des céréales tels que le maïs, et obtenir ainsi un meilleur coût de revient sur la production.

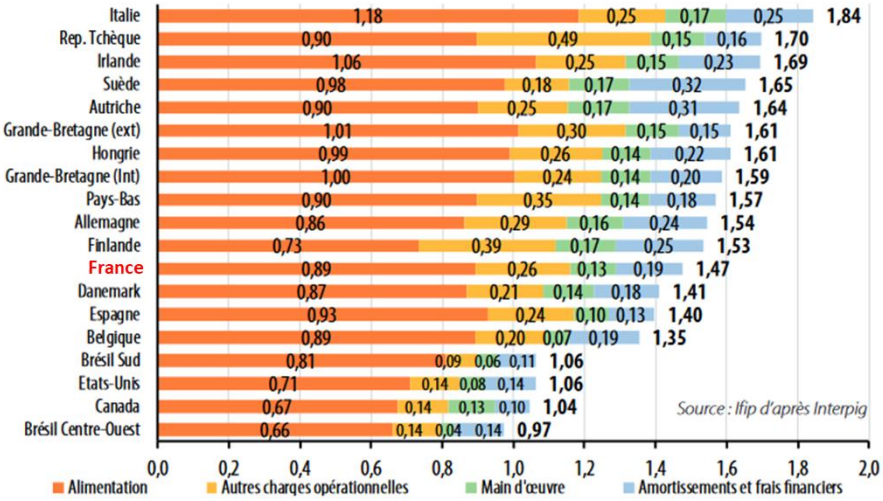

Figure 6 : Coût de revient en euro par kg de carcasse de la filière porcine dans divers pays dont la France (rouge). Les coûts de revient dans chaque pays ont été répartis en 4 grandes catégories : l'alimentation (orange), les autres charges opérationnelles (jaune), la main d'œuvre (vert) et les amortissements et frais financiers (bleu) (IFIP, 2020)

La recherche d'une filière porcine productrice nette de protéines est également favorisée de par le cycle de reproduction court des porcs et leur croissance rapide, et enfin du fait qu'une part plus élevée de leur carcasse entre dans la chaine alimentaire (83% pour un porc) (Figure 7) (Dronne, 2018; GIS, 2017).

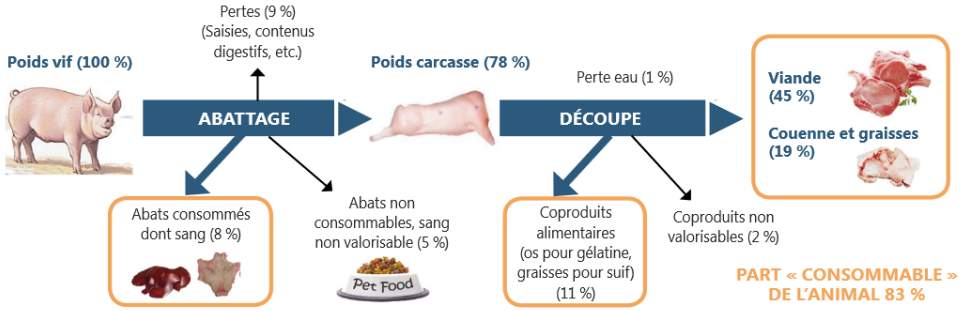

Figure 7 : Exemple de la composition de la carcasse du porc standard (en poids) en partant du poids à vif d'un porc, avec les différents parties du porcs valorisées ou perdues au cours de l'abattage et de la découpe des divers produits issus de la carcasse des porcs (GIS, 2017).

Afin de répondre à une part des questions soulevées par la société, l'amélioration de l'efficacité alimentaire des animaux d'élevages est ainsi apparue comme un moyen de réduire l'impact environnemental pour un indice de consommation similaire à la fois en réduisant l'émission directe d'effluents et en recherchant la valorisation d'une gamme large de coproduits riches en protéines non concurrentielles avec l'alimentation humaine.

2. Les mesures de la consommation et du comportement alimentaire

L'efficacité alimentaire (EA) est aujourd'hui un caractère majeur pour l'économie de nombreuses filières. En production animale, l'EA est définie comme le rapport entre le gain de poids d'un animal en

croissance et le poids d'aliments que l'animal a consommé au cours de la même période. Cette valeur comprise entre 0 et 1 exprime le rendement de transformation de l'aliment par l'animal. Mais classiquement, la mesure utilisée est l'inverse de l'EA, et permet d'exprimer la quantité d'aliment nécessaire pour produire un kg de produit (Indice de Consommation (IC) = Consommation Moyenne Journalière (CMJ) / Gain de poids Moyen Quotidien (GMQ) sur la période évaluée). Depuis plusieurs années de nombreuses études se sont attachées à analyser ce caractère et à déterminer la mesure la plus adaptée pour son évaluation en génétique.

a. *Améliorer l'efficacité alimentaire des animaux*

L'efficacité de l'utilisation des aliments par les animaux est un phénomène très complexe qui dépend de nombreux facteurs physiologiques et environnementaux, néanmoins cinq processus biologiques ont été identifiés comme influençant l'utilisation efficace des aliments par les animaux (Herd and Arthur, 2009). Il s'agit de la capacité d'ingestion, l'utilisation digestive de la ration, l'efficacité métabolique (anabolique et catabolique, notamment associé aux variations de composition corporelle et de développement des viscères et des organes digestifs), la production de chaleur liée notamment à l'alimentation et l'activité physique et enfin la thermorégulation chez les animaux homéothermes. L'importance relative de ces différents processus est variable selon les espèces et les stades physiologiques (croissance, engraissement, gestation, lactation). Ainsi pour le porc le devenir de l'énergie ingérée peut être répartie dans 5 grandes voies présentées sur la figure 8 (Phocas et al., 2014). Parmi les différents leviers d'action, l'amélioration génétique de l'EA afin d'aboutir à une production similaire tout en réduisant l'impact environnemental est apparue comme une solution à explorer.

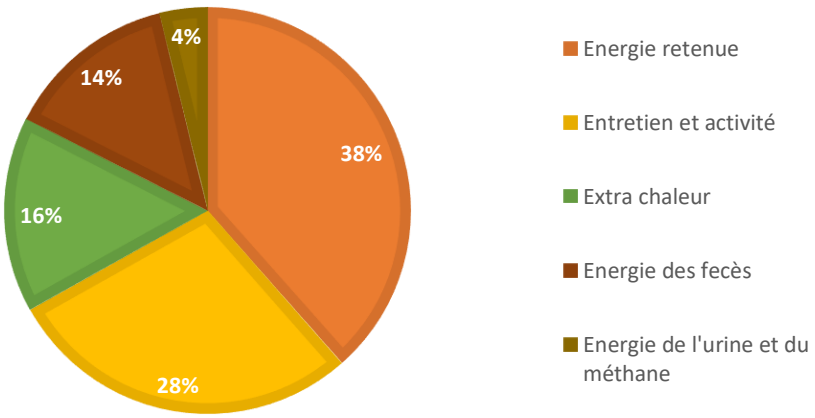

Figure 8 : L'énergie ingérée par les porcs se répercute dans 5 grandes voies : l'énergie emmagasinée, l'énergie utilisée pour leur entretien et l'activité par les individus, l'énergie nécessaire au maintien de la chaleur des individus, l'énergie pour les fèces et l'énergie de l'urine et du méthane (Phocas et al., 2014).

Au vu de la répartition de l'ingéré dans chacun des processus biologiques nécessaires au bon développement des porcs, l'évaluation des besoins alimentaires en fonction de leurs besoins physiologiques est spécifique à chaque individu. Une première étude de l'EA en porc réalisée par Hess et al. (Hess et al., 1941) a permis de démontrer le potentiel d'amélioration de l'EA par sélection sur l'IC du fait de sa variabilité génétique. Cependant la mesure individuelle de l'IC a initialement rarement été utilisée dans les programmes de sélection en raison de la difficulté à estimer la quantité individuelle d'aliments consommés. L'amélioration de l'EA a donc été obtenue indirectement en utilisant des indices de sélection visant à augmenter la vitesse de croissance et réduire la teneur en gras de la carcasse. Cette orientation de la sélection en faveur de porcs à croissance rapide a effectivement permis de réduire l'IC ; depuis les années 80, une forte diminution de l'IC a été observée dans les élevages porcins et est ainsi passée de 3,35 kg d'aliments ingérés par kg de gain de poids en 1984 à 2,73 en 2014 (données IFIP-GTE, 2014) (Gilbert et al., 2017a). Cependant une des réponses défavorables corrélée à l'augmentation de la teneur en muscles des carcasses est simultanément une diminution de l'appétit chez le porc en croissance, avec comme conséquence le risque que les animaux n'expriment plus pleinement leur potentiel de croissance (Labroue, 1995). Ces résultats ont conduit à chercher un autre caractère d'évaluation moins corrélé à la composition corporelle des animaux. En conditions d'élevages classiques, les différences de sexe, de poids et les aléas temporels ou sanitaires auxquels les animaux peuvent être confrontés, expliquent environ 1/3 de la variabilité de l'ingéré chez le porc en croissance. Le second tiers est lié à la vitesse de croissance et la composition corporelle. Pour finir le dernier tiers ne peut être expliqué par des facteurs d'élevages et est qualifié de consommation résiduelle (Gilbert et al., 2017a). C'est à partir de cette décomposition qu'un nouveau critère de l'EA a été proposée afin d'agir sur le levier de la consommation résiduelle par Koch et al. en 1963 (Koch et al., 1963).

b. Le caractère CMJR pour l'efficacité alimentaire

Koch et al. (Koch et al., 1963) ont proposé une autre mesure de l'EA basée sur la différence entre la consommation alimentaire observée et la consommation alimentaire attendue en fonction des besoins de production et d'entretien. En d'autres termes, Koch et al. (Koch et al., 1963) ont suggéré que la consommation alimentaire pouvait être divisée en deux composantes : (1) la consommation d'aliments attendue pour un certain niveau de production et (2) la résiduelle (CMJR pour Consommation Moyenne Journalière Résiduelle). La portion résiduelle de la prise alimentaire peut être en effet utilisée pour identifier les animaux qui s'écartent d'une prise alimentaire correspondant à leurs besoins liés à la croissance et aux besoins métaboliques. Chez les porcs, la consommation d'aliments pour un certain niveau de production représente près de 70 % de la consommation d'aliments, et la CMJR représente les 30 % restants (Cai et al., 2008). Des valeurs de CMJR négatives (ou CMJR-) correspondent à une plus

grande EA parce que l'ingéré observé est inférieur à la valeur prédite. A l'opposé, les animaux moins efficaces présentent une CMJR positive (CMJR+) et consomment une quantité d'aliments supérieure à la quantité estimée pour couvrir les besoins de l'individu (Figure 9). Cette valeur de résiduelle est obtenue par régression linéaire multiple de la CMJ sur la vitesse de croissance estimée par le GMQ, la composition corporelle (via une mesure de l'épaisseur de lard dorsal in vivo (ELD) ou du taux de muscle des pièces à l'abattoir (TMP)), et le poids métabolique moyen (PMM) qui quantifient les besoins d'entretien de l'animal. La consommation résiduelle d'aliments, indépendante des besoins de production et d'entretien, surmonte ainsi les problèmes mentionnés précédemment d'impact négatif de la mesure de l'IC sur les caractères de production des animaux (Kennedy et al., 1993). Cette mesure doit ainsi permettre d'améliorer l'EA tout en maintenant une production constante, voire même en améliorant la production.

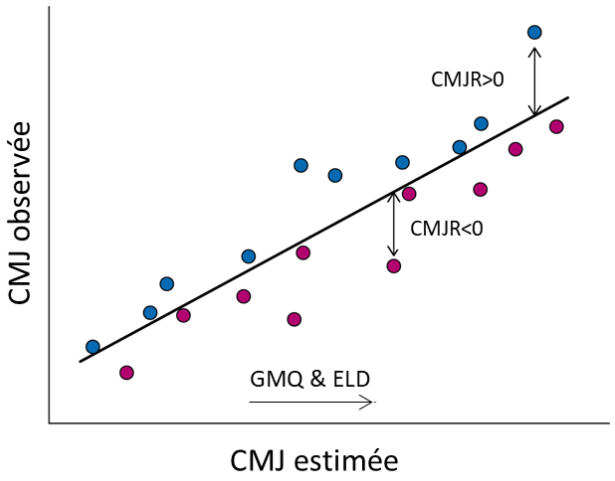

Figure 9 : Schéma de la consommation résiduelle d'aliments comme différence entre la consommation d'aliments observée et la consommation estimée, basée sur le GMQ et l'ELD du porc. L'écart à la droite représente la CMJR : les animaux qui se trouvent au-dessus de la droite ont une CMJR>0 (les moins efficaces, CMJR+), alors que les animaux positionnés en dessous de la droite caractérisent les animaux à CMJR<0 (les plus efficaces, CMJR-) (Azarpajouh et al., 2017).

L'EA des individus peut être améliorée via différentes approches mais pour qu'un caractère phénotypique d'intérêt puisse être candidat à la sélection, celui-ci doit présenter une variabilité génétique, c'est-à-dire que la variabilité de l'expression du phénotype doit avoir une variance génétique additive d'amplitude suffisante. Cette variabilité est estimée à l'aide du calcul de l'héritabilité. Plusieurs études ont indiqué qu'il existait bien une variation génétique pour le CMJR (Table 1) dans les populations porcines de race pure (Johnson et al., 1999; Nguyen et al., 2005; Hoque et al., 2007; Hoque and Suzuki, 2009; Gilbert et al., 2007) et des populations d'animaux croisés (Mrode and Kennedy, 1993; Von Felde et al., 1996). Les valeurs d'héritabilité sont comprises entre 0,22 et 0,47, à l'exception de l'estimation faite par Von Flede et al. (Von Felde et al., 1996), qui était faible (0,18) (Table 1). En 2013, Saintilan et al. (Saintilan et al., 2012) a ainsi analysé 4 races ou lignées élevées et

phénotypées en station de contrôle entre 2000 et 2009, deux lignées femelles (LWF: Large White Femelle et LR: Landrace) et deux lignées mâle (LWM : Large White Mâle et Pi : Pietrain) commerciales. Dans cette étude, la CMJR explique 20,1% en Pi, 26,5% en LWM, 27,6% en LWF, and 29,5% en LR de la variabilité phénotypique de la CMJ (Table 2). Les héritabilités et les corrélations génétiques estimées sont comparables à celles de Cai et al. (Cai et al., 2008) et Gilbert et al. (Gilbert et al., 2007) mesurés sur des dispositifs porcins. Les corrélations génétiques entre la CMJR et les quantités de rejets azotés et phosphorés sont positives pour toutes les races (de 0,51 à 0,58 pour les lignées femelles ; de 0,67 à 0,84 pour les lignées mâles). Saintilan et al. (Saintilan et al., 2012) ont alors proposé différents scenario d'index de sélection dans lesquels l'IC était remplacé par la CMJR et les pondérations sur les autres caractères étaient ajustés en raison de la plus faible corrélation entre ces caractères et la CMJR par rapport à l'IC. Ces valeurs sont comparables à celles estimées pour d'autres espèces animales, comme chez les bovins avec des valeurs comprises entre 0,21 et 0,60 (Crews, 2005; Hardie et al., 2017), ou chez la volaille avec des valeurs comprises entre 0,21 et 0,49 (Aggrey et al., 2010; Begli et al., 2016; Pakdel et al., 2005; Yuan et al., 2015a). Dans leur revue Hoque et Susuki (2009) synthétisent les différentes estimations de paramètres génétiques. Les corrélations génétiques entre la CMJR et les caractères de production sont positives et fortes avec la CMJ (de 0,24 à 0,97 selon les études) et modérées avec le GMQ (0,00 à 0,41 selon les études) (Hoque and Suzuki, 2009). Pour les caractères de carcasse, les corrélations génétiques entre la CMJR et le taux de muscle des animaux est négative quelle que soit l'étude (de -0,31 à -0,61) et de façon cohérente positive avec l'ELD (0,06 à 0,76). Des valeurs de corrélation plus faibles ont été obtenues pour la teneur en gras intramusculaire. Ces résultats indiquent qu'il semble possible de réduire la CMJR en augmentant le Taux de Viande Maigre de la carcasse (TVM) des porcs. Au final, la sélection visant à réduire la CMJR permettrait d'améliorer l’EA et la plupart des phénotypes de carcasse sans compromettre la vitesse de croissance, malgré la réduction de la consommation spontanée d'aliments.

Table 1 : Estimations de l'héritabilité pour la CMJR issues de la littérature *(Hoque and Suzuki, 2009)*

Race[1]	Sexe[2]	# individus	Héritabilité[3]			Source
			$CMJR_1$	$CMJR_2$	$CMJR_3$	
LW	M et F	657	-	-	0,24±0,03	Gilbert et al. (2007)
L	M	341	0,47±0,11	0,29±0,11	-	Hoque and Suzuki (2008)
D	M	514	0,41±0,14	-	-	Hoque et al. (2007a)
LW	M	26 706	0,26±NE*	0,26±NE*	0,26±NE*	Johnson et al. (1999)
Y, L and D	M	7 562	0,33±0,05	0,30±0,06	-	Mrode and Kennedy (1993)
LW	M et F	1 584	0,24±0,08	0,22±0,08	-	Nguyen et al. (2005)
LW and L	M	3 188	-	0,18±0,03	-	von Flede et al. (1996)

Race[1] : LW = Large White ; L = Landrace ; D = Duroc; Y = Yorkshire.
Sexe[2] : M = Mâle ; F = Femelle.
Héritabilité[3] : $CMJR_1$, $CMJR_2$, and $CMJR_3$ = CMJR calculé à partir des modèles incluant le GMQ, ou incluant gain le GMQ et l'ELD, ou encore le GMQ, l'ELD et une mesure de la teneur en muscle, respectivement.
NE* = Non estimé

Néanmoins, la majorité des estimations de paramètres génétiques de la CMJR ont été réalisées durant la période de croissance des animaux dans des conditions d'élevages conventionnelles (alimentation et logement en particulier). Il est donc pour l'heure difficile de déterminer s'il existe ou non une variabilité génétique de la CMJR dans l'ensemble des systèmes de production et la sélection pour ce caractère dans des systèmes diversifiées doivent s'accompagner d'un suivi des réponses corrélées (Hoque and Suzuki, 2009).

Table 2 : Héritabilités(h^2) et corrélations génétiques avec la CMJR (r_g) des caractères mesurés pour les quatre populations (Saintilan et al., 2012)

Population	**LWF**		**LF**		**LWM**		**PP**	
Caractère	h^2	r_g	h^2	r_g	h^2	r_g	h^2	r_g
CMJR (g/j)	0,22	-	0,23	-	0,26	-	0,33	-
CMJ (kg/j)	0,34	0,55	0,27	0,61	0,21	0,72	0,45	0,57
IC (kg/j)	0,30	0,52	0,35	0,56	0,30	0,69	0,36	0,80
GMQ (g/j)	0,33	0,16	0,26	0,05	0,05	0,09	0,42	0,06
RDT (%)	0,39	0,00	0,32	0,03	0,31	-0,09	0,45	-0,05
LARD (mm)	0,51	-0,06	0,62	-0,11	0,53	0,04	0,33	-0,01
TMP (%)	0,61	0,08	0,66	0,03	0,55	-0,04	0,48	-0,12
IQV (point)	0,24	0,18	0,29	0,07	0,12	0,01	0,26	0,23
Nelim (kg)	0,33	0,51	0,30	0,53	0,27	0,67	0,37	0,81
Pelim (kg)	0,28	0,56	0,28	0,58	0,26	0,71	0,37	0,84
A100 (j)	0,31	-0,21	0,36	-0,08	0,38	-0,14	0,40	-0,08
ELD (mm)	0,38	-0,11	0,46	-0,19	0,41	-0,09	0,52	0,01

3. Création de lignées divergentes pour la RFI

L'établissement de lignées divergentes expérimentales est une stratégie pour évaluer les réponses directes et corrélées au caractère de sélection, et pour étudier l'impact de la sélection sur la physiologie des animaux. Deux dispositifs de sélection divergente de porcs sur la CMJR ont été menés : un premier dispositif français sur des porcs Large White de race pure et un second dispositif sur des porcs de race pure Yorkshire à l'ISU aux USA.

a. Les dispositifs expérimentaux de sélection pour la CMJR

i. Dispositif de sélection divergente INRAE

Le dispositif de sélection divergente pour l'efficacité alimentaire sur les porcs Large White (LW) a été initié par Pierre Sellier et Jean Noblet en 1999, puis conduit par Gilbert et al. (Gilbert et al., 2007) depuis les années 2000 pendant 10 générations de sélection (lignées INRAE) à partir de 30 couples d'animaux F0 permettant de représenter la diversité de la population LW. A chaque génération, 96 verrats issus de la génération précédente ont été testés pour le caractère CMJR et 6 mâles ont été sélectionnés par lignée et accouplés avec 35 à 40 femelles non sélectionnées de la même génération et la même lignée.

A chaque génération plusieurs parités ont été produites afin d'abord de tester et sélectionner les males reproducteurs de la génération suivante, et ensuite de disposer d'animaux contrôles (femelles et mâles castrés) pour estimer les réponses directes et corrélées de la sélection sur différents caractères (Figure 10) (Gilbert et al., 2017a).

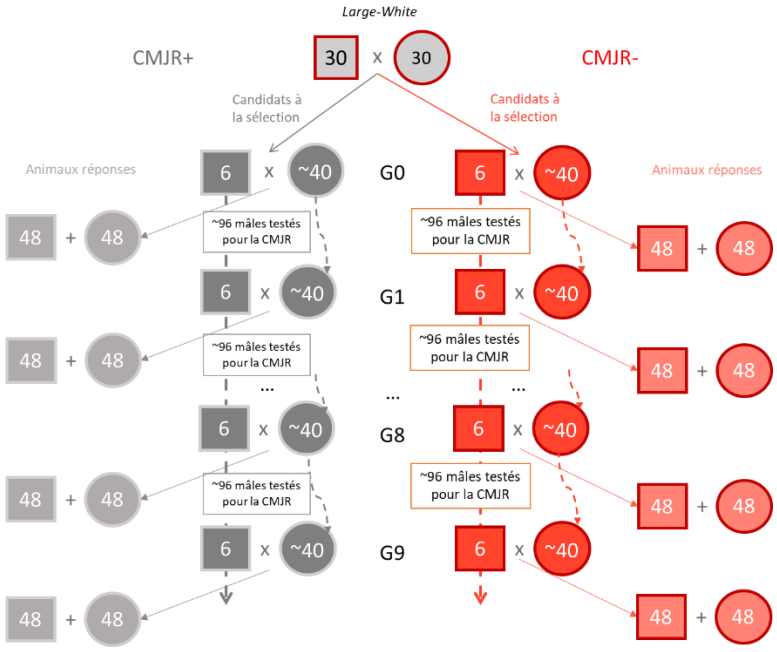

Figure 10 : Schéma de la sélection divergente sur le caractère CMJR des porcs Large White (Lignées INRAE) avec la création de 2 parités (les candidats à la sélection et les animaux réponses) pour chacune des lignées divergentes (CMJR+ et CMJR-). Cette sélection divergente a été réalisée sur 10 générations, de la G0 à la G9. La P1 se caractérise par l'accouplement des 6 mâles sélectionnés à la génération précédente (parmi 96 mâles testés) avec environ 40 femelles (non testées et remplacées par une fille à la génération suivante). La P2 se compose d'environ 48 mâles et 48 femelles issus de l'accouplement des mêmes parents.

Pour le choix des verrats reproducteurs, la consommation alimentaire des animaux a été enregistrée via des distributeurs automatiques de concentrés (DAC) ACEMA 64 au cours de la période comprise entre 35 à 95 kg de poids corporel. Le critère de sélection pour la CMJR a été calculé en utilisant les enregistrements individuels des verrats testés pour la CMJ, le GMQ et l'ELD à 95 kg selon l'équation : $CMJR_1,g = CMJ_1,g - (1,06 \times GMQ_1,g) - (37 \times ELD_1,mm)$ (Gilbert et al., 2007). Pour les porcs utilisés pour estimer la réponse à la sélection, toute la période de croissance a été prise en compte pour le test, qui a duré du troisième jour après l'entrée des porcs dans les loges équipées de DAC (moyenne du poids vif (PV0) = 28 kg) jusqu'à la veille de l'abattage (moyenne PV1 = 107 kg). La valeur de la CMJR pour les animaux réponses est calculée comme suit : $CMJR_2, g = CMJ_2, g - (1,48 \times GMQ_2, g) + (23,2 \times TVM,\%) - (99,1 \times PMM,kg)$, où le poids métabolique moyen (PMM) correspond à $(PV1^{1,6} - PV0^{1,6})/[1,6 (PV1 - PV0)]$. Pour chaque modèle, les effets fixes de la bande de contrôle, du sexe (pour $CMJR_2$) et de la taille du groupe testé ont été ajoutés dans le modèle. Une héritabilité de 0,13 pour l'index de sélection a été estimée. Pour les 23 caractères complémentaires mesurés sur les animaux

réponses, les héritabilités estimées par Gilbert et al. (Gilbert et al., 2017b) s'échelonnent de 0,04 pour le temps d'imbibition de la viande jusqu'à 0,59 pour la teneur en viande maigre (Table 3).

Table 3 : Paramètres génétiques estimés (h^2=héritabilité ; ρ_g = corrélation génétique avec la CMJR ; σ_g = écart-type génétique ; σ_p = écart-type phénotypique), et réponses génétiques à la sélection en génération G9 dans les lignées CMJR+ et CMJR-, ainsi que le niveau de signification de la différence (p-value) : p-value *pour la différence entre les moyennes des moindres carrés des valeurs génétiques pour les lignées CMJR- et CMJR+ de la génération G9 avec *** = P<0,001, * = 0,01<P<0,05, t = 0,05<P<0,10. (Gilbert et al., 2017b)*

Caractères	h^2	ρ_g	σ_g	σ_p	CMJR+	CMJR-	*p-value*
Index de sélection	0.13		6.63	18.60	11.00 (0.27)	−14.10 (0.27)	***
CMJR	0.13		42.93	119.56	73.88 (1.71)	−91.03 (1.71)	***
IC	0.42	0.39	0.13	0.20	0.15 (0.01)	−0.17 (0.01)	***
CMJ	0.41	0.25	127.63	199.85	146.22 (6.18)	−123.67 (6.18)	***
GMQ	0.50	−0.07	54.12	76.45	10.53 (2.71)	2.31 (2.71)	*
Rendement de carcasse (RDT)	0.36	0.05	1.06	1.76	−0.50 (0.05)	0.49 (0.05)	***
Poids de longe	0.54	0.15	0.42	0.57	−0.38 (0.02)	0.33 (0.02)	***
Poids de jambon	0.51	0.09	0.32	0.45	−0.24 (0.02)	0.07 (0.02)	***
Poids d'épaule	0.38	0.06	0.28	0.45	−0.15 (0.01)	0.16 (0.01)	***
Poids du lard dorsal	0.43	−0.08	0.30	0.46	0.13 (0.02)	−0.14 (0.02)	***
Poids de poitrine	0.28	0.11	0.21	0.39	0.23 (0.01)	−0.17 (0.01)	***
TVM	0.59	0.14	2.02	2.64	−1.72 (0.10)	0.92 (0.10)	***
ELD	0.40	0.02	2.40	3.80	−0.23 (0.12)	−0.32 (0.12)	
pH 24h adducteur (AD)	0.41	0.28	17.06	26.57	14.69 (0.86)	−9.01 (0.86)	***
pH 24h muscle semi membraneux (SM)	0.38	0.22	12.03	19.48	12.90 (0.60)	−8.08 (0.60)	***
pH 24h fessier superficiel (GS)	0.39	0.23	10.84	17.31	12.36 (0.53)	−9.08 (0.53)	***
pH 24h longe dorsale (LM)	0.32	0.19	10.41	18.50	9.99 (0.48)	−5.13 (0.48)	***
L* muscle fessier moyen (GM)	0.20	−0.14	1.59	3.55	−0.24 (0.07)	0.36 (0.07)	***
L* muscle fessier superficiel (GS)	0.33	−0.17	2.13	3.71	−2.27 (0.09)	1.83 (0.09)	***
a* fessier moyen (GM)	0.29	−0.09	1.26	2.34	0.05 (0.06)	−0.43 (0.06)	***
a* fessier superficiel (GS)	0.26	0.12	0.88	1.73	0.09 (0.04)	−0.02 (0.04)	t
b* fessier moyen (GM)	0.24	−0.12	0.80	1.65	−0.02 (0.04)	−0.09 (0.04)	
b* fessier superficiel (GS)	0.32	−0.08	0.84	1.48	−0.55 (0.04)	0.41 (0.04)	***
temps d'imbibition du muscle	0.04	−0.29	10.11	48.22	3.80 (0.32)	−3.11 (0.32)	***
IQV	0.33	0.26	1.61	2.80	1.97 (0.08)	−1.44 (0.08)	***

Colorimétrie L*=brillance, a*=rouge, b*=jaune, mesuré 24 h après l'abattage

ii. Dispositif de sélection divergente de l'Iowa State University (ISU, USA)

En parallèle des lignées divergentes sélectionnées en France, un second dispositif a été mis en place à l'Iowa State University (ISU, Iowa, USA). Ce dispositif de sélection divergente a été mené par Jack Dekkers depuis 2000 à partir d'animaux de race Yorkshire pure (lignée ISU) (Cai et al., 2008).

Dans un premier temps, de la G0 à la G4, une seule lignée a été sélectionnée, correspondant aux animaux les plus efficaces (lignée CMJR-), et en parallèle une lignée contrôle a été maintenue par croisement au hasard (Cai et al., 2008). A chaque génération, la CMJ, le GMQ et l'ELD ont été mesurés pendant une période de test comprise entre environ 40 kg et 115 kg sur 90 verrats d'une première parité et 90 femelles produites dans une seconde parité de la lignée CMJR-. Après évaluation des verrats de première parité, environ 12 mâles et 70 femelles CMJR- de la première parité étaient sélectionnés pour produire environ 50 portées à la génération suivante. La sélection a été réalisée sur

la valeur génétique de la CMJR estimée à partir de la CMJ, le GMQ et l'ELD. Les données enregistrées sur la seconde parité étaient utilisées pour consolider les prédictions de valeurs génétiques de la génération suivante. Un résumé du protocole de sélection de la CMJR- est présenté sur la figure 11a. En parallèle, environ 30 portées de la lignée contrôle ont été produites par accouplements aléatoires entre 10 mâles et 40 femelles (G0 à G4) (Figure 11b). Dans un second temps, à partir de la génération G5, la sélection d'animaux moins efficaces (CMJR+) a débuté dans la lignée dite contrôle (Young et al., 2011) : 12 verrats ont ainsi été sélectionnés parmi 90 candidats à la sélection, et 70 cochettes parmi environ 250 candidates sœurs des verrats ont été sélectionnées sur la base de valeurs génétiques estimées à partir des informations familiales, pour produire environ 50 portées pour la génération suivante (Figure 11b).

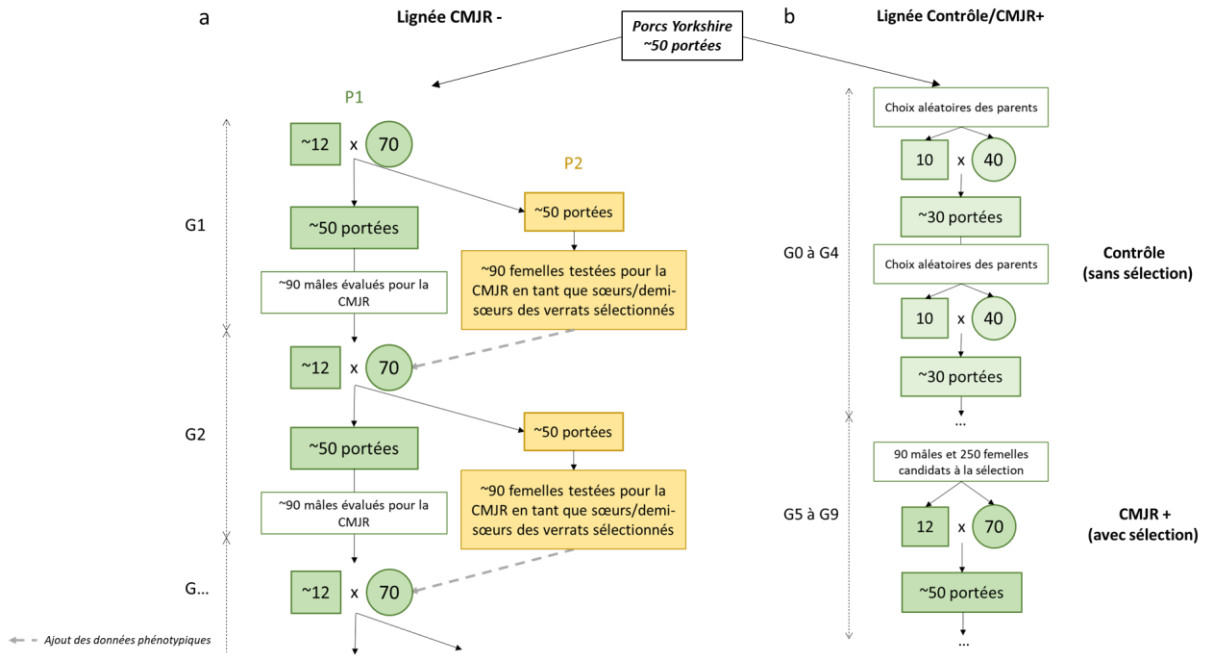

Figure 11 : Schéma du dispositif de sélection des porcs Yorkshire, (a) avec la création de la lignée CMJR- (Lignées ISU) et la création de deux parités (P1 en vert ; P2 en orange) caractérisées pour la P1 par les mâles testés et sélectionnés en même temps que les femelles non testées par rapport à leurs valeurs génétiques pour la CMJR, et pour la P2 par l'accumulation de données phénotypiques pour la CMJR de sœurs ou demi-sœurs des reproducteurs sélectionnés. (b) La lignée contrôle a été obtenue par sélection aléatoire des reproducteurs jusqu'à la génération G4 (vert clair) puis une lignée CMJR+ a été sélectionnée sur base de leurs valeurs génétiques pour la CMJR (vert).

La deuxième parité a été produite à chaque génération en répétant les accouplements de la parité 1 pour recueillir des enregistrements phénotypiques supplémentaires sur environ 90 cochettes par lignée. La sélection s'est poursuivie de cette manière pendant 10 générations. Au final, 7 caractères ont été étudiés et Cai et al. (Cai et al., 2008) ont estimé l'héritabilité de ces caractères. Le caractère CMJR possède une héritabilité de 0,29 et la plus faible héritabilité a été obtenue avec le caractère

d'efficacité alimentaire (0,17) alors que le caractère d'épaisseur de lard dorsale possède la plus forte héritabilité (0,68) (Table 4).

Table 4 : Paramètres génétiques estimés pour le dispositif des lignées ISU. RFI = CMJR ; ADFI = CMJ ; ADG = GMQ ; FE = efficacité alimentaire (EA) ; BF = ELD ; LMA = superficie de la longe ; IMF = teneur en gras intramusculaire. *(Cai et al., 2008)*

Trait[2]	n	Mean	SD[1]	Heritability	Litter[3]	Pen(group)[3]	Residual
RFI, g/d	756	0	126	0.29 ± 0.07	0.01 ± 0.00	0.30 ± 0.06	0.40 ± 0.07
ADFI, g/d	756	1,989	216	0.51 ± 0.08	0.00 ± 0.00	0.13 ± 0.04	0.36 ± 0.08
ADG, g/d	756	768	91	0.42 ± 0.08	0.00 ± 0.00	0.02 ± 0.02	0.56 ± 0.08
FE, %	756	38.76	3.30	0.17 ± 0.07	0.05 ± 0.00	0.16 ± 0.04	0.62 ± 0.07
BF, mm	756	15.88	3.48	0.68 ± 0.09	0.08 ± 0.01	0.00 ± 0.00	0.24 ± 0.09
LMA, cm^2	756	42.67	4.67	0.57 ± 0.10	0.11 ± 0.01	0.02 ± 0.02	0.30 ± 0.09
IMF, %	492	1.75	0.40	0.28 ± 0.11	0.27 ± 0.02	0.01 ± 0.02	0.44 ± 0.10

b. Impact de la sélection divergente sur les autres caractères des animaux sélectionnés

i. Sur les caractères d'efficacité alimentaire et de composition corporelle

Les réponses directes et corrélées à la sélection ont été évaluées en génération G5 (Gilbert et al., 2009) puis en génération G9 (Gilbert et al., 2017b) pour les lignées INRAE ; les résultats obtenus sont résumés dans la figure 12 et la table 2. Après neuf générations de sélection, la sélection divergente pour la CMJR a conduit à des différences entre lignées très significatives de ce caractère (-165 g/jour dans la lignée CMJR- par rapport à la lignée CMJR+ (P<0,001)), de la CMJ (-270 g/jour) et une réponse marquée sur l'IC (-0,32 kg d'aliments/kg de gain, P<0,001). De faibles réponses ont été observées pour la vitesse de croissance (-12,8 g/jour, P<0,05) et la composition corporelle (+0,9 mm de l'ELD, P=0,57 ; -2,64 % de TVM, P<0,001) entre les lignées. Les réponses à la sélection des lignées ISU sont semblables (Cai et al., 2008). En effet, Cai et al. (Cai et al., 2008) indiquent également que la lignée CMJR- possède, par rapport à la lignée contrôle, une différence significative en terme de CMJR (−96 ± 28 g/j, P< 0,0001), une CMJ inférieure (−165 ± 35 g/j, P< 0,0001), un IC inférieur (P< 0,0001), un GMQ inférieur (−33 ± 14 g/j, P= 0,022) et un ELD inférieur (−1,99 ± 0,76 mm, P= 0,013) (Figure 13 et table 3). Ils ont aussi noté un rapport gain de poids/kg d'aliments ingérés supérieur (efficacité alimentaire, 1,36 ± 0,78 % P= 0,09) dans la lignée CMJR- comparée aux animaux à la lignée témoin. Les deux dispositifs (INRAE et ISU) ont également démontré que les porcs CMJR- ont des besoins énergétiques d'entretien plus faibles (P<0,02 pour les lignées INRAE et P<0,13 pour les lignées ISU) que les porcs CMJR+ (Barea et al., 2010; Boddicker et al., 2011).

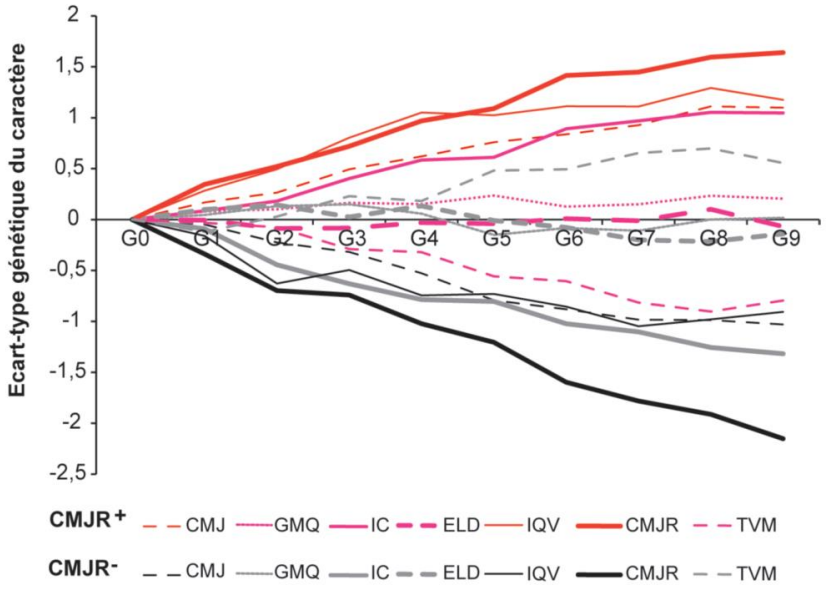

Figure 12 : Réponses à la sélection sur la Consommation Moyenne Journalière Résiduelle (CMJR) pour les caractères de croissance, de consommation et de composition de carcasse de la génération G0 à la génération G9, exprimées en écart-types génétiques des caractères. CMJR+ = lignée à CMJR élevée, moins efficace ; CMJR– = lignée à CMJR faible, plus efficace ; CMJ = consommation moyenne journalière ; GMQ = Gain Moyen Quotidien ; IC = Indice de Consommation ; ELD = Épaisseur de Lard Dorsal ; IQV = Indice de Qualité de la Viande ; TVM = Taux de Viande Maigre de la carcasse estimé par combinaison linéaire des poids de morceaux (Gilbert et al., 2017a).

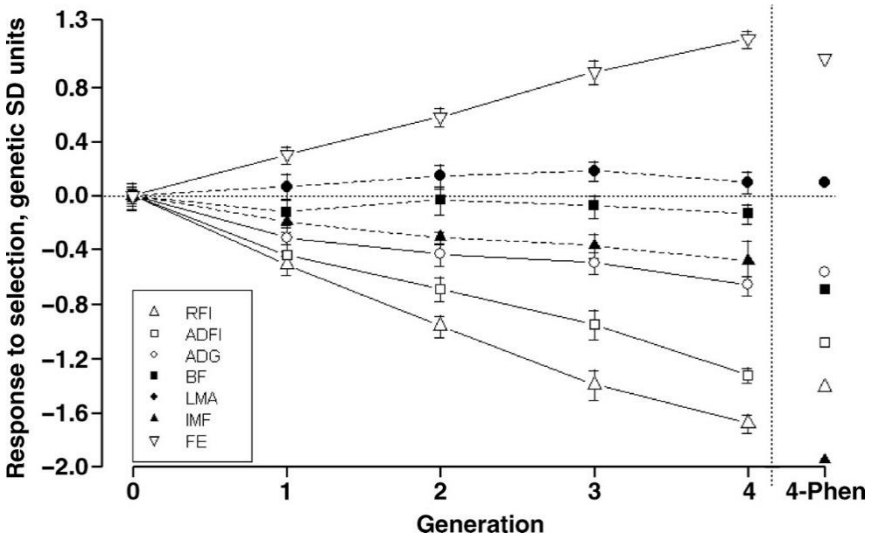

Figure 13 : Réponses directes et corrélées à la sélection sur la consommation alimentaire résiduelle (CMJR) basées sur la valeur génétique moyenne et la comparaison directe des lignées en génération 4 (4-Phen). Les valeurs génétiques moyennes et les barres d'erreurs standards sont basées uniquement sur les verrats CMJR-. 4-Phen : Différences phénotypiques des lignées (contrôle et CMJR-) basées sur la comparaison phénotypique des lignées de la génération 4 des lignées ISU ; ADFI = consommation moyenne journalière ; FE (efficacité alimentaire) = ADG/ADFI ; BF = épaisseur de lard dorsal ; LMA = surface de muscle de la longe ; IMF = teneur en gras intramusculaire (Cai et al., 2008).

ii. Sur la qualité de viande

La qualité de viande a également été influencée par la sélection pour la CMJR. À partir de la troisième génération de sélection divergente pour la CMJR des lignées INRAE, les caractères de qualité technologiques et les indicateurs sensoriels de la qualité de la viande ont été altérés dans la lignée CMJR- tant au niveau de la longe que dans le jambon (Gilbert et al., 2017b). Un pH plus faible et une augmentation de la clarté (L*) de la viande (P<0,001) ont été observés chez les animaux CMJR- du dispositif INRAE sans impact sur la qualité sensorielle de la viande (Faure et al., 2013). Ces différences de qualité de la viande ont été attribuées à la présence d'un plus grand nombre de fibres glycolytiques de plus gros diamètre dans la longe, pour les animaux CMJR- (P < 0,02) (Lefaucheur et al., 2011). Par contre, les différences de pH ou de capacité de rétention d'eau entre la lignée CMJR- et la lignée témoin n'étaient pas significatives pour les lignées ISU jusqu'à la G4 (Cai et al., 2008). Mais dans les générations suivantes, Arkfled et al. (Arkfeld et al., 2015) ont signalé des différences de qualité entre les lignées, avec notamment une perte en eau, des différences de couleurs et une teneur en eau plus élevée chez les porcs CMJR-. La qualité et la composition de la viande fraîche ont été évaluées chez les porcs G5 par Smith et al. (Smith et al., 2011) pour la lignée ISU, et les résultats ont montré que les porcs de la lignée CMJR- avaient une surface de longe supérieure (P< 0,05), une plus grande quantité de viande maigre (P < 0,05) et moins de gras intramusculaire (P < 0,01) par rapport aux porcs de la lignée témoin. Au final, la qualité de la viande a été impactée négativement par la sélection pour une réduction de la CMJR dans les deux dispositifs, mais ces différences n'ont eu qu'un impact mineur sur la qualité sensorielle de la viande fraiche.

iii. Sur le comportement alimentaire

Des différences dans le comportement alimentaire ont été observées dans les lignées INRAE et l'activité des porcs a aussi été impactée dans les deux dispositifs de sélection (INRAE et ISU). En G6 pour les lignées INRAE, les porcs CMJR- ont consommé moins d'aliments (P < 0,001), ont réalisé moins de visites aux DAC (P < 0,001) mais ont consommé plus d'aliments par visite, engendrant une vitesse d'ingestion plus élevée (P < 0,001) que les porcs CMJR+ (Gilbert et al., 2017b). En revanche, pour la lignée ISU sélectionnée, en générations G4 et G5, plusieurs caractéristiques du comportement alimentaire ont également été mesurées mais aucune différence n'a été détectée entre les porcs CMJR- et les porcs de la lignée contrôle pour la consommation d'aliments par visite ou par heure, ni pour le nombre de visites au DAC par jour ou par heure (Young et al., 2011). Par contre, les deux dispositifs présentent des différences d'activités entre les porcs CMJR- et les porcs contrôles ou CMJR+. Les porcs CMJR- ont passé moins de temps debout, moins de temps au DAC et étaient globalement moins actifs (Gilbert et al., 2017b; Meunier-Salaün et al., 2014; Sadler et al., 2011).

iv. Pour la digestibilité

Aucune différence notable n'a été établie entre les lignées françaises pour la digestibilité de l'aliment ingéré avec un régime alimentaire standard (à base de céréales et de farine de soja). De ce fait, la digestibilité des aliments ne semble pas jouer un rôle important dans l'explication de la variation de la CMJR dans les lignées (Gilbert et al., 2017b). Les porcs CMJR- des lignées ISU nourris avec une alimentation américaine conventionnelle avaient une meilleure efficacité de digestion dans les premières générations (Harris et al., 2012), mais n'a pu être retrouvé qu'avec une alimentation riche en fibres et pauvre en énergie dans les générations suivantes (Mauch et al., 2018).

v. Sur les processus métaboliques des porcs

Le métabolisme est un processus physique clé qui contribue à la variation de l'efficacité alimentaire (Herd and Arthur, 2009). Ainsi, plusieurs processus métaboliques ont été étudiés mais seul le métabolisme énergétique présentait des différences cohérentes entre essais entre les lignées INRAE. En G6, les porcs CMJR- avaient un métabolisme énergétique plus faible que les porcs CMJR+, comme l'indiquent des activités glycolytiques et une oxydation des acides gras β plus faibles. Dans les lignées américaines, le muscle des porcs CMJR- avaient tendance à avoir une activité de la calpaïne inférieure ($P \leq 0,10$), une activité de la calpastatine supérieure ($P < 0,05$), un rapport calpaïne:calpastatine inférieur ($P < 0,05$) et une activité du protéasome inférieure ($P < 0,05$) par rapport au tissu musculaire des porcs CMJR+. De plus, les porcs CMJR- de l'ISU avaient une dégradation protéique moindre que les porcs CMJR+, ce qui pourrait contribuer à une meilleure efficacité alimentaire des animaux CMJR-, car moins d'énergie serait nécessaire pour la dégradation des protéines dans cette lignée (Cruzen et al., 2013). En complément de ces résultats, les études des profils protéiques mitochondriaux pour les deux dispositifs ont indiqué que les porcs CMJR- peuvent présenter un stress oxydatif moindre ce qui suggèrerait que les porcs CMJR- pourraient disposer de plus d'énergie orientée pour leur croissance que les porcs CMJR+ (Grubbs et al., 2013; Vincent et al., 2015). Néanmoins pour les lignées INRAE, lors de l'analyse du métabolisme des porcs des deux lignées divergentes, le taux de renouvellement des protéines ne présentait aucune différence (Barea et al., 2010; Labussière et al., 2015), et l'activité mitochondriale était similaire entre les lignées dans la longe.

Pour finir, les porcs CMJR- INRAE présentaient des concentrations plus faibles du facteur de croissance analogue à l'insuline I (IGF-I) en post-sevrage (environ -18 ng/ml), qui est un polypeptide présent dans le sang circulant et associé à la croissance et au développement postnatals, par rapport aux porcs CMJR+ ($P< 0,001$) (Gilbert et al., 2017b). Dans les lignées ISU en G5, l'IGF-I a également été retrouvé significativement plus faible (-47 ng/ml) dans la lignée CMJR- par rapport à la lignée témoin ($P < 0,0001$) (Bunter et al., 2010).

4. Vers la prédiction génomique de l'efficacité alimentaire

En complément des dispositifs de sélection divergente pour étudier les bases génétiques et biologiques de ce caractère d'efficacité alimentaire, d'autres études de la CMJR ont été menées sur des dispositifs issus des populations commerciales.

La prédiction génomique consiste à prédire pour un caractère d'intérêt la valeur génétique des animaux à partir de l'information de génotypage pour un grand nombre de marqueurs génétiques positionnés sur le génome. Ces estimations sont classiquement obtenues en remplaçant la matrice de parenté pedigree dans le modèle linéaire mixte de prédiction des valeurs génétiques par une matrice d'apparentement obtenue d'après les informations génomiques. Les effets des génotypes aux marqueurs peuvent alors être estimés à partir d'individus génotypés et qui ont des enregistrements de performance (population de référence). Ces estimations pour le caractère d'intérêt sont alors utilisés dans la population des animaux génotypés qui n'ont pas de phénotype pour estimer une valeur génomique au caractère (GEBV : Genomic Estimated Breeding Value) (Meuwissen et al., 2001). Une limite majeure des approches de prédiction génomique est la précision d'estimation des valeurs dans de petites populations dont les populations de référence sont de taille limitée. Néanmoins, aujourd'hui de nouvelles méthodologies ont été développées et rendues robustes, comme par exemple les méthodes single-step GBLUP (ssGBLUP) qui combinent les informations phénotypiques, génotypiques et généalogiques dans une seule évaluation génomique (Aguilar et al., 2010; Christensen and Lund, 2010; Legarra et al., 2009), qui permettent de mieux exploiter l'intégralité de l'information disponible dans une population.

La précision des prédictions génomiques pour les caractères de production, dont l'efficacité alimentaire chez les porcs en croissance a été étudiée dans plusieurs races porcines. En effet, du fait du coût de la mesure du phénotype automatiquement sur des animaux en groupe, l'efficacité alimentaire est un des caractères de production les plus coûteux à obtenir, et qui est disponible sur un sous-échantillon des candidats à la sélection seulement, ce qui affecte la précision des prédictions. La prédiction génomique pourrait permettre de contrecarrer cette limite. Une première étude a été menée sur le GMQ et l'IC sur des porcs Duroc danois (Christensen et al., 2012). Des données génomiques obtenues à partir de la puce Illumina Porcine SNP60 BeadChip étaient disponibles pour environ 2 700 porcs mesurés pour le GMQ dont 1 500 disposaient également un IC. Par rapport à une prédiction d'après les informations pedigree, les précisions des prédictions obtenues pour le GMQ et l'IC en utilisant une méthode ssGBLUP passaient de 0,18 à 0,35 et de 0,18 à 0,21 respectivement. Une troisième étude a confirmé certains de ces résultats à partir d'une population d'environ 1 400 porcs Duroc (Zhang et al., 2018), mesurés pour le GMQ, la CMJ ainsi que l'ELD et l'épaisseur de longe. Les précisions moyennes des prédictions génomiques pour ces caractères étaient respectivement de 0,15,

0,40, 0,65 et 0,30. L'ingestion étant un caractère coûteux à mesurer, il existe néanmoins peu de littérature sur les prédictions génomiques pour l'efficacité alimentaire et les caractéristiques de ses composants chez les porcs (Table 5).

Table 5 : Table récapitulative des précisions de prédictions des caractères liés à l'efficacité alimentaire calculées sur des populations de porcs. La corrélation entre les prédictions génomiques et le phénotype a été divisée par la racine carrée de l'héritabilité et certaines valeurs (b) ont été arrondies à partir de la fiabilité rapportée dans les publications. (Zhang et al., 2018)

Trait	Accuracy	Breed and reference
Days to 250 lbs	0.66–0.84	Yorkshire (Badke et al., 2014)
ADG	0.50–0.58[b]	Danish Duroc (Ostersen et al., 2011)
	0.40–0.43[b]	Danish Duroc (Guo et al., 2016)
	0.24	Duroc (Jiao et al., 2014)
FCR	0.39–0.45[b]	Danish Duroc (Ostersen et al., 2011)
	0.11	Duroc (Jiao et al., 2014)
FAT	0.69–0.86	Yorkshire (Badke et al., 2014)
	0.55–0.56[b]	Danish Duroc (Guo et al., 2016)
	0.37	Duroc (Jiao et al., 2014)
ADFI	0.15	Duroc (Jiao et al., 2014)
RFI	0.09	
LMD	0.30	

Trait = Caractère ; Accuracy = Prédiction ; Breed and reference = Race et référence bibliographique ; ADFI = CMJR ; FAT = ELD ; ADG = GMQ ; LMD = épaisseur du muscle lombaire.

Quelques études récentes rapportent l'évaluation génomique du caractère CMJR. Jiao et al. (Jiao et al., 2014) sont un des premiers à avoir évalué ce caractère chez le porc, en complément des caractères précédemment discutés. Leur analyse a été réalisée sur une population d'un peu plus de 1 000 porcs Duroc pour lesquels des phénotypes tels que le poids des individus à différents stades de croissance, des mesures d'épaisseurs de lard et de muscle ou encore d'enregistrements individuels d'aliments ingérés ont été récoltés. En utilisant environ 1 000 animaux comme référence et 500 animaux pour la validation, les précisions des prédictions génomiques estimées (Table 5) variaient de 0,094 à 0,365, la valeur la plus faible étant obtenue pour la CMJR. La valeur d'héritabilité de la CMJR étant plus faible par rapport aux autres caractères étudiés, les auteurs suggèrent que le dispositif est certainement trop petit pour atteindre une précision élevée.

Afin d'apporter un autre regard sur la prédiction génomique pour des caractères d'efficacité alimentaire et notamment la CMJR, Aliakbari et al. (Aliakbari et al., 2020) ont réalisé des études de prédiction génomiques en utilisant des animaux issus de la sélection divergente pour la CMJR (Gilbert et al., 2017b). Cette étude a permis d'évaluer le potentiel de construction d'un jeu de données d'entrainement (population de référence) pour la prédiction génomique des caractères d'efficacité alimentaire à partir d'individus plus ou moins apparentés issus de lignées sélectionnées connectées génétiquement et de petite taille. Ainsi, Aliakbari et al. (Aliakbari et al., 2020) ont montré que la prédiction génomique utilisant une population de référence incluant des animaux issus de lignées apparentées et sélectionnées pourrait être aussi précise que la prédiction génomique obtenus à partir d'une population de référence construite intra-lignée. Les précisions ainsi obtenues allaient de 0,28 à 0,66 à effort de génotypage comparable entre scénarios. Cette stratégie peut être particulièrement favorable pour l'estimation des caractères qui sont fortement affectés par la sélection en raison d'une plus forte variabilité représentée au sein de la population de référence, mais les auteurs rapportent un risque de biais dans les prédictions pour ces caractères.

II. Cartographie des régions QTL pour l'efficacité alimentaire

1. Structure de la variabilité du génome de porc et cartographie

a. Evolution des outils d'analyse de la variabilité du génome

L'architecture génétique des caractères quantitatifs est complexe et l'identification des gènes qui sous-tendent la variation génétique nécessite un grand nombre de marqueurs génétiques. Les microsatellites sont ainsi devenus l'un des marqueurs génétiques les plus populaires au cours des années 1990 (Litt and Luty, 1989). La distribution homogène des microsatellites le long du génome, leurs fréquences et leurs hauts niveaux de polymorphisme ont grandement facilité la construction des cartes génétiques (Archibald et al., 1995; Ellegren et al., 1994; Rohrer et al., 1994). Bien que les cartes de liaison par microsatellite avec des intervalles de 5-10 cM soient suffisantes pour la primo localisation des loci ayant des effets importants sur un phénotype (Lipkin et al., 1998), la caractérisation des QTL ayant des effets plus faibles nécessite des cartes génétiques de résolution plus élevées. A la fin des années 90, l'évolution des techniques de séquençage (séquençage de troisième génération) a permis l'obtention de la première séquence du génome humain (International Human Genome Sequencing Consortium, 2001; Venter et al., 2001). Par la suite, sur le même modèle, de nombreux consortiums internationaux ont été mis en place pour les espèces modèles ou les espèces d'intérêt agronomique. En 2003, le séquençage des 18 autosomes et des 2 chromosomes sexuels du génome porcin a été initié par la création d'un consortium (SGSC) composé de représentants des milieux universitaires, gouvernementaux et industriels. (Archibald et al., 2010; Schook et al., 2005). Une première version de la séquence du génome porcin (DraftV9) a été publiée en 2009 (Humphray et al., 2007). En parallèle des projets complémentaires de séquençage partiel d'individus de différentes races ont permis l'identification d'un grand nombre de variants ponctuels de séquence (polymorphismes). En 2009, Ramos et al. (Ramos et al., 2009) ont entrepris une recherche systématique de SNP chez le porc afin de produire une première puce de génotypage de marqueurs SNP. Cette puce a été créée à partir de 5 mélanges d'échantillons ADN comprenant de 23 à 36 individus, représentant les quatre races principalement utilisées en production porcine (Duroc, Piétrain, Landrace et Large White) et un mélange d'individus de différentes races pour intégrer des variants issus d'ADN de sangliers. Près de 373 000 variants ont été identifiés et ajoutés aux SNPs d'ores et déjà identifiés et publiés dans plusieurs bases de données. A partir de cet ensemble de 510 000 SNPs, une sélection basée sur, la fréquence des allèles mineurs (MAF), le nombre d'allèles, l'espacement entre marqueurs et la localisation des SNPs sur les chromosomes, a été appliquée afin de concevoir une puce contenant 64 232 SNPs avec un espacement moyen de 43,4 kb entre les SNPs. La validation ultérieure par le génotypage de 554 échantillons porcins représentant diverses races a été réalisée et a permis de valider un call rate (CR) moyen supérieur à 0,99 (Illumina, 2015). Depuis la publication de la première séquence du génome

porcin, deux mises à jours ont été réalisées. En 2012, grâce à l'ajout de séquences réalisées à partir de librairies « whole genome shotgun », WGS, et l'utilisation de séquences d'ADNc permettant l'annotation du génome Groenen et al. (Groenen et al., 2012) ont proposé une nouvelle version de la séquence de référence porcine (Draft V10.2). Plus récemment une nouvelle version du génome de porc (Draft V11.1) a été publiée (Warr et al., 2020). Les améliorations du génome de porc présentent dans la version 11.1 sont liées à l'utilisation de méthodes de séquençage plus récentes que pour la version précédente, permettant ainsi de diminuer la présence de séquences redondantes, le nombre d'erreurs d'ordre et d'orientation de contig de séquences, et aussi d'augmenter la qualité de l'assemblage. En parallèle de nouvelles versions de puces de génotypage de SNP ont été développées : une seconde puce moyenne densité, comprenant 40K SNP en commun avec la précédente a été développée par Illumina. L'objectif du développement de cette nouvelle version, GGP Porcine HD array (GeneSeek, Écosse, Royaume-Uni), comprenant plus de 68 516 SNPs, était la prise en compte des nouvelles connaissances sur le génome porcin et possède ainsi une meilleure répartition des variants le long des chromosomes porcins. Enfin, une puce très haute densité comprenant 658 692 marqueurs répartis sur l'ensemble des chromosomes et les marqueurs les plus représentatifs des puces moyennes densités a été mise au point (Groenen, 2015). L'ensemble des puces porcines ont été reportées dans la table 6. L'avantage majeure de ces nouveaux outils de génotypage est la prise en compte de manière plus précise du déséquilibre de liaison présent au sein des races porcines.

Table 6 : Etat de l'art des différentes puces de génotypage porcines présentent sur le marché.

Nom de la puce	Nombre de SNPs	Entreprise	Technologie
PorcineSNP60 BeadChip v2 array	64 232	Illumina	Illumina Infinium chemistry
GGP Porcine HD Array	68 516	Illumina	Illumina Infinium chemistry
Axiom® Genome-Wide Pig genotyping Array (high-density panel)	658 692	Affymetrix	Axiom assay

b. Importance du déséquilibre de liaison pour la cartographie

Le déséquilibre de liaison (DL) se définit comme une association non aléatoire d'allèles à différents loci. Dans le cas de deux loci bi-alléliques (SNP A et SNP B) avec respectivement des allèles A/a et B/b, le déséquilibre de liaison reflète une fréquence d'association des allèles aux deux loci, *p*AB, *p*Ab, *p*aB et *p*ab supérieure (ou inférieure) aux fréquences attendues compte-tenu des fréquences de chaque allèle

*p*A, *p*a, *p*B et *p*b: D = pAB – pA*pB = pab – pa*pb (Lewontin, 1964). La valeur *D* est une mesure quantitative de l'association allélique qui dépend fortement des fréquences alléliques. Deux mesures normalisées de la valeur *D* ont donc été proposées afin de corriger pour ce biais (Ardlie et al., 2002). Le *D'*, proposé par Lewontin (Lewontin, 1964) :

$$D' = \frac{D}{D_{max}}, \text{ avec } D_{max} = \begin{cases} \min(pAb; papB) \text{ si} D > 0 \\ \min(pab; pAB) \text{ si} D < 0 \end{cases}$$

Ou encore la valeur du r^2 qui représente le coefficient de corrélation entre allèles (Hill and Robertson, 1966) : $r^2 = \frac{D^2}{pApapBpb}$. Cette dernière mesure du DL est moins sensible aux fréquences alléliques extrêmes et est également moins sensible à la taille de la population étudiée (Weiss and Clark, 2002), c'est pourquoi elle est souvent privilégiée dans les études de populations animales basées sur des SNPs du fait de sa robustesse. Cette mesure prend des valeurs comprises entre 0 et 1, où 1 correspond à un cas de DL total.

Le projet Human HapMap (Gibbs et al., 2003; Thorisson et al., 2005) a révélé une grande variation du DL dans le génome humain (Reich et al., 2001). Ce projet a également permis de décrire la présence de différences importantes de DL parmi les populations humaines, qui résultent de différences dans l'histoire et la démographie des populations (Ardlie et al., 2002; Reich et al., 2001). Néanmoins Reich et al. (Reich et al., 2001) ont estimé un DL moyen présentant un r^2 d'environ 0,1. Pour les animaux, le DL a été étudié chez diverses espèces d'élevage, par exemple les bovins (Porto-Neto et al., 2014) où les valeurs de DL à courte distance (< 10 kb) étaient comprises entre un r^2 de 0,25 et 0,45 pour deux races bovines de type Nelore et Angus respectivement. Pour les ovins (McRae et al., 2002), les porcs (Nsengimana et al., 2004) et les poulets (Aerts et al., 2007) les valeurs de r^2 étaient toutes trois d'environ 0,3 sur une distance de 1 cM en moyenne. En raison de la petite taille efficace des populations et d'une forte sélection artificielle, les animaux d'élevage ont un DL bien supérieur à celui retrouvé chez l'homme et plusieurs études du DL ont été menées afin de le caractériser dans diverses races porcines. Amaral et al. (Amaral et al., 2008) ont étudié l'étendue du DL chez des porcs de races européennes et chinoises, ainsi que chez le sanglier européen, dans 3 régions génomiques sur les chromosomes 3 et 18. Les 3 régions génomiques sélectionnées disposent d'une forte densité en marqueurs SNP. Dans cette étude, ils ont mis en évidence une décroissance du DL qui diffère fortement entre les races porcines chinoises et européennes (Figure 14). Au final, le DL est principalement organisé en blocs de 10 kb maximum chez les races de porcs asiatiques, tandis que chez les races européennes, les blocs de DL peuvent atteindre 400 kb (Amaral et al., 2008). Nsengimana et al. (Nsengimana et al., 2004) ont évalué le DL dans cinq populations de porcs commerciaux (Large White, Landrace, Duroc/Large White et Yorkshire/Large White) dans deux régions chromosomiques de 40 cM,

l'une sur le SCC4 avec un *D'* qui variait de 0,150 à 0,215 et l'autre sur le SCC7 où le *D'* variait de 0,208 à 0,340. Cette différence entre région du génome est justifiée, par les auteurs, par la présence de nombreux QTL dans la région du SSC7 associés au taux de croissance et au dépôt de graisse chez le porc. Nsengimana et al. (Nsengimana et al., 2004) suggèrent que les différences en terme de taille de DL seraient ainsi dues à une pression de sélection différente entre ces régions.

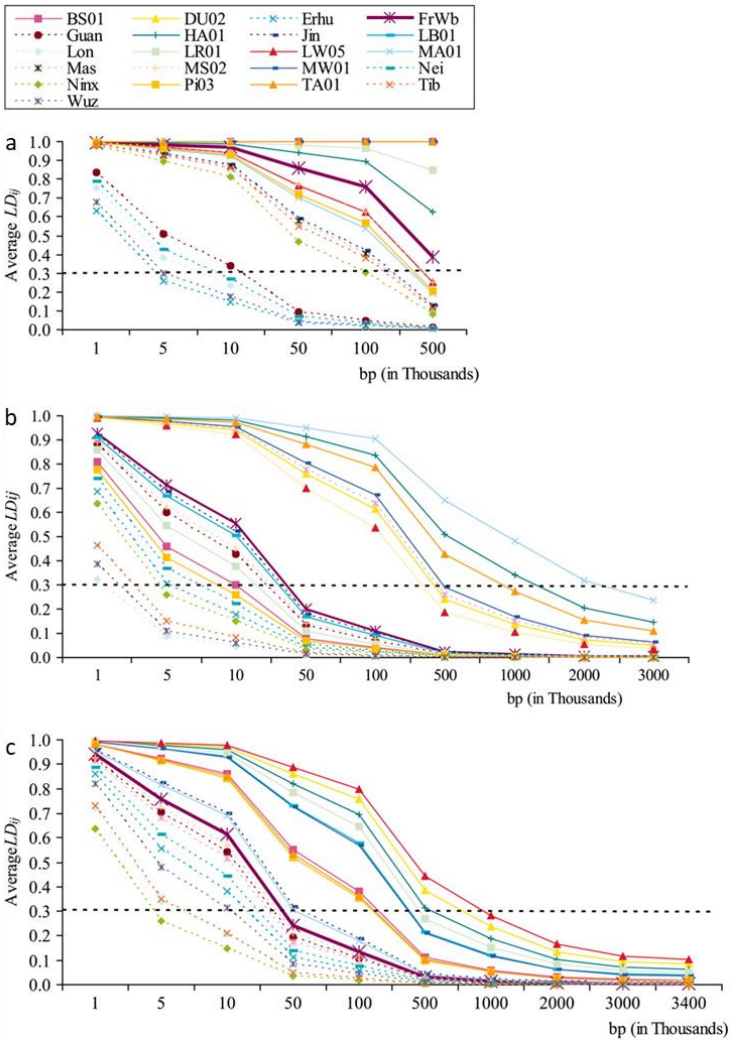

Figure 14 : Le déséquilibre de liaison (DL) moyen pour toutes les populations à différentes distances présentées en kb. La relation entre le DL prédit (LD_{ij}) et la distance (paires de bases) est indiquée par race et par région génomique : cible 1 (a), cible 2 (b) et cible 3 (c). Les races chinoises sont représentées par des lignes pointillées, les races européennes par des lignes pleines, et le sanglier par une ligne pleine épaisse (Amaral et al., 2008).

La connaissance du DL dans une population d'étude est extrêmement précieuse pour localiser les gènes affectant les caractéristiques quantitatives, identifier les régions chromosomiques en cours de sélection, étudier l'histoire des populations et gérer les ressources et la diversité génétiques. De plus, le DL se maintient d'autant plus dans le temps, et au fil des générations, que deux loci sont proches et donc liés (Ardlie et al., 2002). Les puces de SNP développées sont destinées à capturer le DL

dans la majorité des populations ; plus la distance entre marqueurs de la puce diminue et plus le DL capturé entre les marqueurs sera grand. Badke et al. (Badke et al., 2014) ont ainsi évalué la mesure du DL dans 4 populations en fonction de la densité des marqueurs disponibles et ont conclu que la décroissance du DL était relativement faible lorsque la distance entre marqueurs augmentait jusqu'à 1 Mb (Figure 15). Le DL élevé à des distances de 1Mb ou plus est un indicateur que les puces permettant l'analyse simultanée d'au moins 50 000 SNPs peut grandement faciliter la recherche de loci influençant la variabilité des caractères via des analyses d'association.

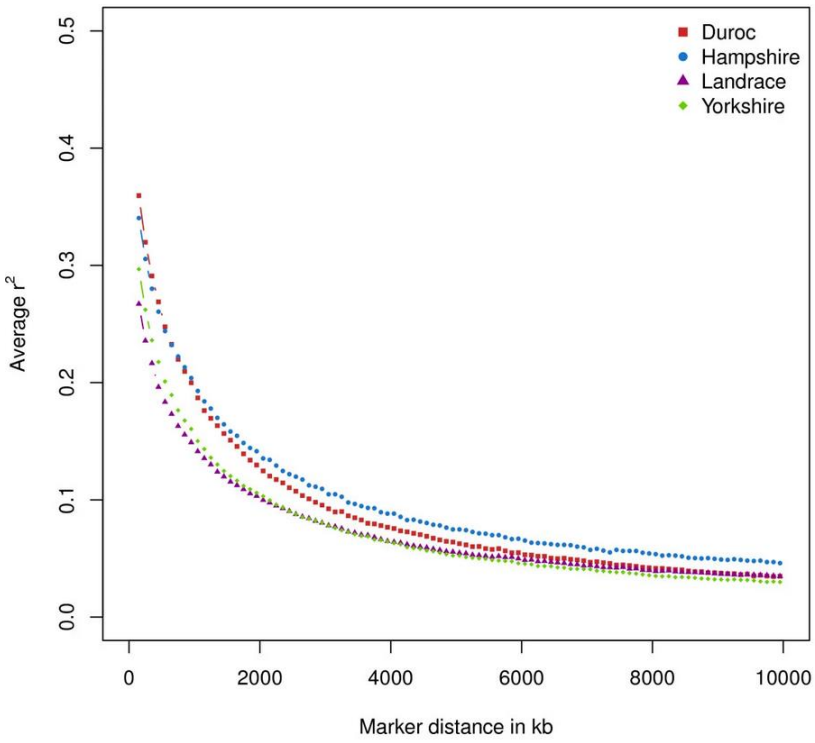

Figure 15 : Distribution du déséquilibre de liaison via le r^2 de 0 à 10 Mb dans 4 populations de porcs Américaines (Badke et al., 2014)

c. Les études d'associations, outil clé d'identification de régions QTL

Pour la recherche de mutations sous-jacentes aux QTL (Quantitative Trait Nucleotide, QTN), la stratégie la plus commune aujourd'hui consiste à utiliser des études d'association tout génome (« *Genome Wide Association Studies* », GWAS). Le principe consiste à génotyper de nombreux individus, mesurés également pour le caractère d'intérêt, sur un nombre de marqueurs très élevés. Le but des GWAS est donc d'identifier les associations génotype-phénotype. Pour cela de multiples algorithmes ont été développés pour augmenter à la fois l'efficacité de calcul et la puissance statistique des modèles linéaires mixtes. Aujourd'hui, différents logiciels sont à notre disposition comme Plink (Purcell et al., 2007), EMMAX (Kang et al., 2010), GCTA (Yang et al., 2011) ou GEMMA (Zhou and Stephens, 2012).

GEMMA s'adapte à un modèle linéaire mixte univarié (LMM) pour les tests d'association sur un seul phénotype afin de tenir compte de la stratification de la population et de la structure de l'échantillon, et pour estimer la proportion de variance des phénotypes expliqués par les génotypes (Zhou and Stephens, 2012).

Les GWAS peuvent être utilisées soit sur des cohortes d'individus dits cas-témoins pour tester des associations avec des maladies, soit des cohortes destinées à identifier des associations pour des caractères quantitatifs. Dans les deux cas, on suppose que les cohortes sont composées d'individus non apparentés qui partagent le même milieu populationnel, bien que cela puisse ne pas être le cas en pratique. La présence d'individus apparentés au sein d'un échantillon d'étude donne lieu à une structuration de l'échantillonnage, et ce terme englobe à la fois les problèmes de stratification de la population et les problèmes d'apparentés cachés (Kang et al., 2010). La stratification de la population fait référence à l'inclusion d'individus de différentes populations au sein d'un même échantillon d'étude, alors que la parenté cachée fait référence à la présence de relations génétiques inconnues entre des individus au sein de l'échantillon d'étude. Les effets de structure de l'échantillon présents dans les cohortes utilisées pour les études d'association génétique ont été bien documentés et identifiés comme étant la cause de fausses associations (Helgason et al., 2005).

Les GWAS ont révolutionné le domaine de la génétique des maladies complexes au cours de la dernière décennie, en fournissant de nombreuses régions QTL pour des maladies et des caractères complexes (McCarthy et al., 2008). La première publication basée sur les GWAS date de 2002. Ozaki et al. (Ozaki et al., 2002) ont effectué des analyses GWAS sur l'infarctus du myocarde. D'autres études d'associations pour différentes maladies ont rapidement suivies, Klein et al. (Klein et al., 2005) ont identifié 2 SNPs liés à la dégénérescence maculaire liée à l'âge, Speakman et al. (Speakman et al., 2018) ont publié des travaux sur l'obésité, et Watanabe et al. (Watanabe et al., 2018) sur le diabète de type 2. Fin 2010, plus de 3 000 articles rapportant des études GWAS avaient été publiés jusqu'à atteindre en 2019 plus de 5 000 publications, toutes espèces confondues (Figure 16a). Chez les espèces d'élevage, suite à la commercialisation des puces de génotypage pour les espèces bovine, porcine, ovine, caprine et aviaire, une augmentation croissante et régulière du nombre d'analyses GWAS publiées est également constatée (Figure 16b). Chez le porc des études ont été menées pour les différents caractères quantitatifs importants pour la filière, comme l'étude de Duijvesteijn et al. (Duijvesteijn et al., 2010) qui ont identifiés, à l'aide d'études d'associations, plusieurs régions du génome responsables de la variation des niveaux d'androstérones chez les verrats, l'étude du comportement alimentaire des porcs (Do et al., 2013), ou encore Schneider et al (Schneider et al., 2012) qui ont réalisé des études d'associations sur des caractères de mise bas chez les porcs. L'augmentation croissante des études GWAS a induit une augmentation du nombre de régions QTL

identifiées. À ce jour (Décembre 2020) rien que pour l'espèce porcine, 31 000 QTL sont référencés dans la base de données PigQTLdb (Hu et al., 2019).

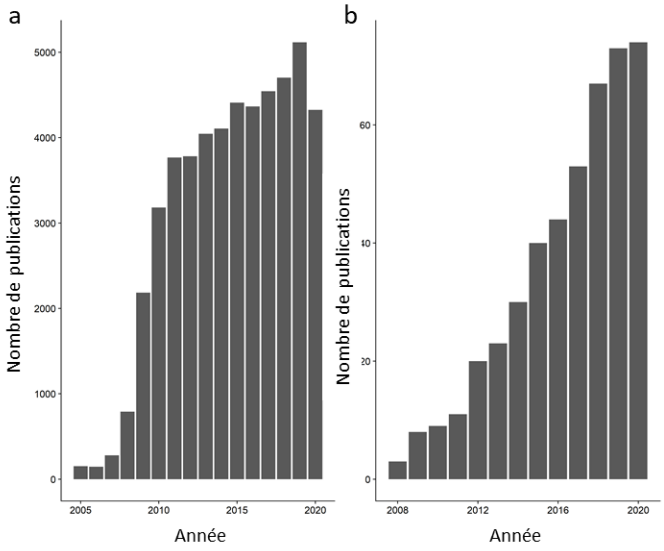

Figure 16 : Nombre de publications par an utilisant les méthodes GWAS pour tous types d'organismes vivants (a) et sur les GWAS pour les animaux d'élevages (b) (source : pubmed Décembre 2020)

Un point important de la cartographie fine de régions QTL est l'utilisation de génotypes les plus denses possibles permettant une bonne représentation de la variabilité, afin de détecter un signal d'association avec les QTN recherchés (Korte and Farlow, 2013) (Figure 17). S'il est possible d'utiliser une puce haute-densité pour obtenir cette densité de génotypage, la méthode généralement utilisée pour y parvenir, est le recourt à l'imputation.

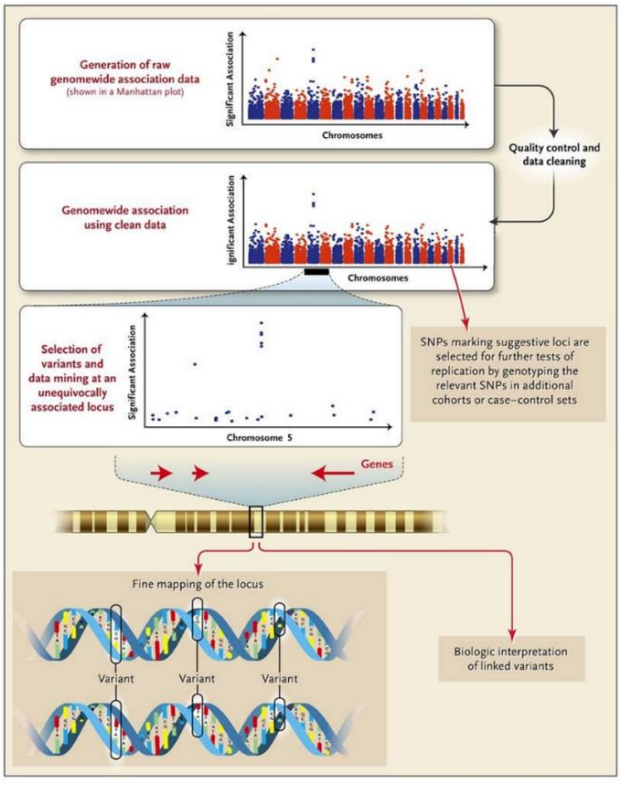

Figure 17 : Schéma illustrant les analyses réalisées des études GWAS jusqu'à l'identification de QTN sur le génome ciblé

2. L'imputation de génotypes pour affiner la localisation de variants

a. Principe

Le principe de l'imputation est d'utiliser une population de référence disposant de génotypes complets obtenus à partir d'une puce à ADN de moyenne ou haute densité selon l'étude (Figure 18a), pour compléter les génotypes manquants d'une population étudiée pour un nombre de marqueurs plus faible (Figure 18b). Même si la population candidate possède une densité faible de marqueurs, elle doit obligatoirement posséder un lot de SNPs communs avec les SNPs du panel de référence. Les méthodes d'imputation tentent ainsi d'identifier le partage d'haplotypes entre les individus de référence et les individus à imputer (Figure 18c), et utilisent ce partage de génotypes pour combler l'information manquante chez les individus candidats (Excoffier and Slatkin, 1995; Stephens et al., 2001). L'hypothèse émise dans la plupart des modèles d'imputation, est que les haplotypes d'un individu donné peuvent être modélisés comme une mosaïque d'haplotypes connus dans la population de référence (Figure 18d) (Marchini and Howie, 2010). Les facteurs pris en compte dans la réalisation de l'imputation sont la densité de marqueurs dans la population à imputer et dans la population servant de base à l'imputation, ainsi que la fréquence allélique des marqueurs dans la population (Li et al., 2009). Le principe de l'imputation peut être appliquée pour des génotypes obtenus à partir de

puces de génotypage de densité différente, mais également entre des puces de génotypage et la séquence complète du génome d'un lot d'individus. Dans le cas de l'imputation jusqu'à la séquence, le processus est similaire ; un noyau d'individus est entièrement séquencé, et les individus restants génotypés sur une puce haute densité (HD) seront imputés jusqu'à la séquence en utilisant les individus séquencés comme référence (van Binsbergen et al., 2014).

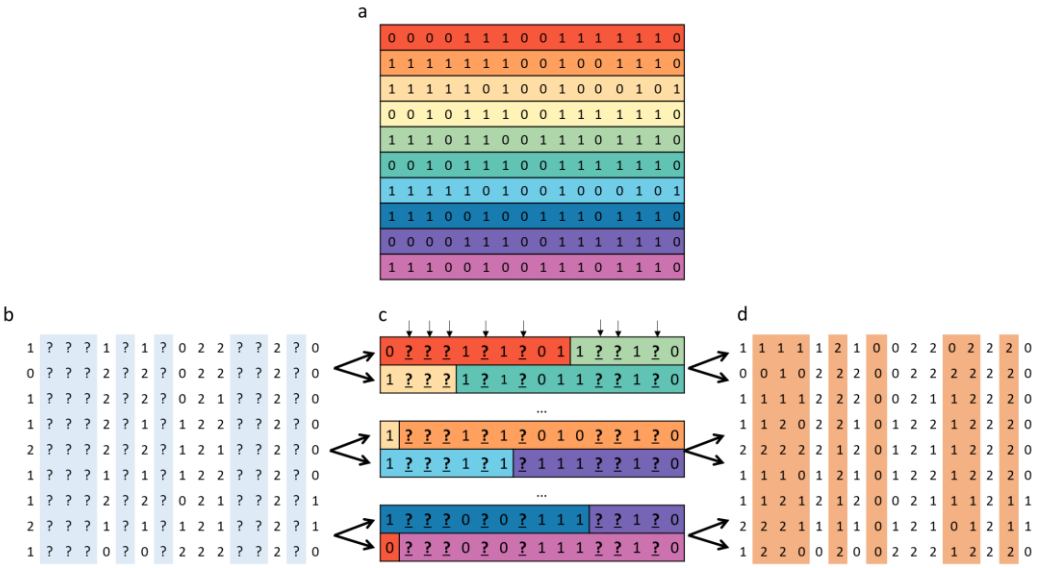

Figure 18 : Schéma explicatif du processus d'imputation, avec (a) des individus possédant un panel de génotypes complets et (b) des individus avec des données génotypiques manquantes. (c) La combinaison de la population de référence et de la population candidate permet de retrouver les haplotypes correspondants. Les haplotypes identifiés dans la population candidate à l'imputation, ont été colorés en fonction des haplotypes de référence auxquels ils correspondent. Au final, (d) la population candidate obtient un génotype complet après imputation. (Marchini and Howie, 2010)

En 2009, Li et al. (Li et al., 2009) ont mis en évidence l'importance de la taille de la population dans les approches d'imputation (Figure 19). La qualité des analyses fondées sur l'imputation augmente à mesure que la taille des panels de référence augmente. Cette augmentation de précision est due au fait que les haplotypes partagés entre les échantillons de l'étude et les échantillons du panel de référence augmentent en taille et sont donc plus faciles à identifier sans ambiguïté par rapport au panel de référence plus grand.

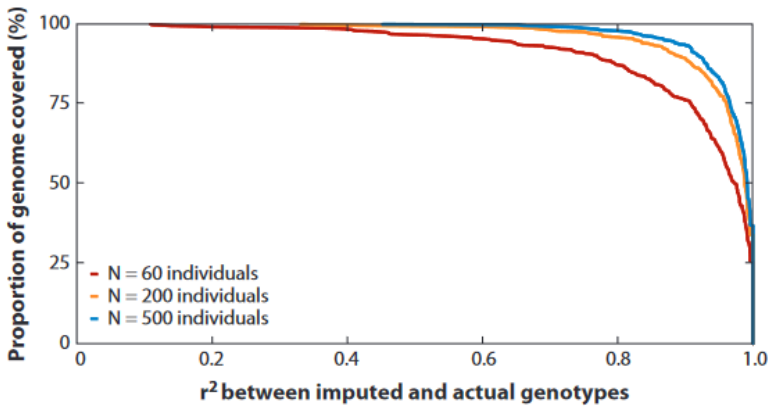

Figure 19 : Représentation de la qualité d'imputation selon différentes tailles de population, N=60 individus (rouge), N=200 individus (orange) ou N=500 individus (bleu). Les courbes illustrent l'impact de la proportion de marqueurs dont les génotypes sont imputés avec précision (r^2 élevé entre les génotypes imputés et réels). (Li et al., 2009)

Les logiciels pour imputer les génotypes manquants d'une population candidate sont assez nombreux. Les logiciels Beagle (Browning and Browning, 2009) et IMPUTE2 (Howie et al., 2009), développés initialement pour des applications en génétique humaine, utilisent un modèle de Markov caché (HMM) pour déduire les marqueurs manquants. La méthode d'imputation de Beagle construit un arbre d'haplotypes et le résume dans un graphique acyclique direct en joignant les nœuds de l'arbre en fonction de la similarité des haplotypes. La méthode d'imputation IMPUTE2 est basé sur l'estimation alternative des haplotypes dans le panel de référence et l'imputation des génotypes manquants dans le panel de test en choisissant les haplotypes les plus similaires. Le logiciel FImpute (Sargolzaei et al., 2014) est lui basé sur l'utilisation des informations issues du pedigree, car les parents proches partagent généralement des haplotypes plus longs, tandis que les parents plus éloignés partagent des haplotypes plus courts. La relation entre les animaux testés et les animaux de référence a une influence considérable sur la précision de l'imputation comme l'a également démontré Jonhson et al. (Johnston et al., 2011). Jonhson et al. ont indiqué que les logiciels testés basés sur l'exploitation de l'information familiale (FImpute, AlphaImpute (Hickey et al., 2012), findhap (VanRaden et al., 2011) et PHASEBOOK (Druet et al., 2010)) ont eu une proportion de génotypes correctement imputés nettement meilleure lorsque les relations entre les ancêtres génotypés et les animaux à imputer sont proches. L'impact du degré d'apparentement entre les populations candidates et de référence a également été rapporté par Ma et al. (Ma et al., 2013) qui ont comparé 6 logiciels d'imputation sur deux populations de bovins. D'après leur étude, les outils IMPUTE2 et Beagle sont des méthodes plus précises et robustes pour tirer parti des informations familiales. Ils ont également montré que FImpute était l'outil le plus adapté pour imputer des génotypes MD (50K) vers une densité HD (500K) en raison non seulement de la précision de l'imputation mais également du temps de calcul plus faible qu'avec les autres logiciels.

De plus, Manolio et al. (Manolio et al., 2009) ont indiqué que les variants génétiques à faibles fréquences pouvaient jouer un rôle très important dans le déterminisme des caractères complexes et pouvaient également avoir des effets plus importants que les variants communs. Par conséquent, l'efficacité de l'imputation des marqueurs avec une faible MAF est un facteur important pour évaluer la qualité d'imputation des logiciels, et de ce fait Ma et al. (Ma et al., 2013) ont réalisé les calculs des taux de génotypes corrects et des coefficients de corrélation pour chacun des logiciels testés. Leurs résultats montrent que si le taux de génotypes corrects était plus élevé lorsque les MAF étaient plus faibles (Figure 20a), en revanche le coefficient de corrélation était plus faible (Figure 20b).

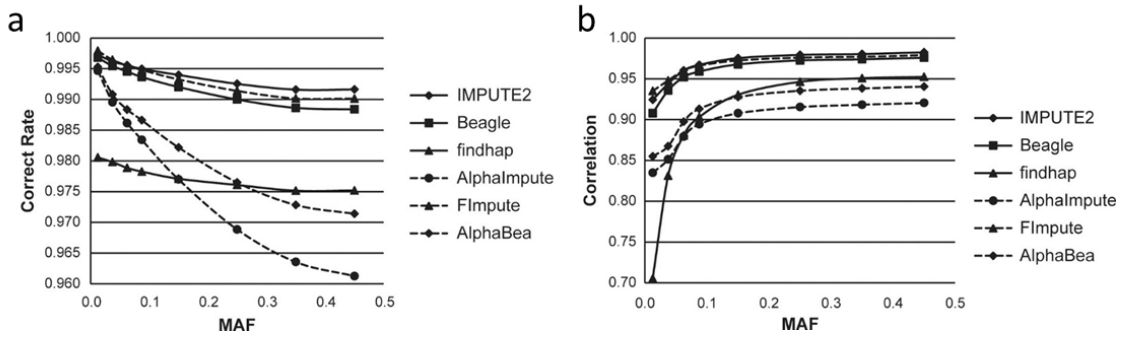

Figure 20 : (a) Taux de génotype correct par rapport aux fréquences et (b) corrélation entre les valeurs des vrais génotypes et des génotypes imputés par rapport aux fréquences alléliques des SNPs pour l'imputation moyenne densité (54 000 marqueurs) à haute densité (777 000 marqueurs) en utilisant différentes méthodes d'imputation : IMPUTE2, Beagle, findhap, AlphaImpute et FImpute. AlphaBea signifie que les génotypes manquants après imputation par AlphaImpute ont été imputés par Beagle.(Ma et al., 2013)

b. *Les différents domaines d'utilisation de l'imputation*

Les approches d'imputation ont été appliquées à deux domaines d'étude distincts : les études d'association et la prédiction génomique. Pour les analyses GWAS, l'analyse de marqueurs génotypés et de marqueurs imputés (afin d'obtenir une densité plus forte) peut conduire à la détection d'associations significatives et à une vue plus détaillée des régions associées (Spencer et al., 2009). Le facteur déterminant est alors le choix de la densité de marqueurs à utiliser respectivement dans la population de référence et la population imputée afin de garantir la qualité des génotypes imputés. Chez le porc plusieurs études ont été publiées sur l'utilisation de l'imputation pour obtenir des données génétiques à haute densité. Dans l'ensemble des études, la précision de l'estimation des génotypes imputés est très élevée avec des valeurs de corrélation s'échelonnant de 0,88 (Badke et al., 2014) à 0,99 (Gualdrón Duarte et al., 2013) pour des niveaux d'imputation de faibles densités (<10K) jusqu'à un niveau moyenne densité (~60K). La plus faible qualité d'imputation a été obtenue par Badke et al. (Badke et al., 2014) qui rapporte une précision moyenne d'imputation de 0,88 en utilisant un petit panel de référence d'haplotypes (environ 130 haplotypes). Néanmoins, cette précision d'imputation a été largement améliorée, avec une corrélation de 0,95, lorsqu'un panel de référence plus important (environ 1 800 haplotypes) a été utilisé. Ces résultats sont cohérents avec ceux de Gualdrón Duarte et

al. (Gualdrón Duarte et al., 2013) qui ont utilisé une population F2 génotypée avec 9 000 SNPs pour les imputer jusqu'à une densité de 60K en utilisant les génotypes des individus F0 et F1 génotypés avec la puce 60K.

Si la majorité des études GWAS sur données imputées a été réalisée à partir de génotypes HD imputés, ces dernières années, l'imputation des génotypes à l'échelle de la séquence complète a été prisée. L'un des avantages de l'utilisation des données de séquence du génome entier par rapport aux génotypes HD pour les études GWAS est que les mutations à l'origine des différences phénotypiques sont incluses dans les données utilisées pour l'analyse (Daetwyler et al., 2014). Comme le QTN est inclus, le déséquilibre de liaison entre un SNP analysé et le QTN n'est plus limitant. Par conséquent, parmi les tests d'associations réalisés, l'association la plus significative sera obtenue avec le QTN (van Binsbergen et al., 2014). Chez les animaux domestiques, la majorité des GWAS réalisés sur individus dont le génome a été imputé a pour l'heure était principalement réalisée chez les bovins (Daetwyler et al., 2014; Pausch et al., 2017a). Chez le porc, Zhang et al. (Zhang et al., 2018) pour une population de Duroc et Van den Berg et al. (van den Berg et al., 2019) pour des porcs Large White, ont obtenu une précision d'imputation d'environ 0,92 et 0,99 respectivement, pour une imputation d'une densité de marqueurs de 80K à 650K. Pour l'imputation jusqu'à la séquence, Van den Berg et al. (van den Berg et al., 2019) ont défini des précisions d'imputation dans deux populations de porcs, des Landrace Allemand et des Large White, en utilisant un panel de référence multi-populations. La qualité d'imputation jusqu'à la séquence pour les porcs Large White possédant des génotypes haute densité (660K) a été de 0,93. Par contre, les auteurs ont aussi reporté que l'imputation jusqu'à la séquence des génotypes des porcs Landrace, qui ne disposaient que de génotypes moyenne densité (80K) présentaient une qualité d'imputation moins bonne avec un score de 0,84.

Le second domaine d'application des approches d'imputation est la prédiction génomique. Le génotypage des candidats à la sélection sur des puces MD ou HD peut ne pas être rentable lorsque le nombre de candidats à la sélection est élevé ou que l'avantage économique par candidat à la sélection est faible par rapport au coût du génotypage (Hayes et al., 2012), comme chez le porc. Mais les coûts du génotypage sont considérablement plus faibles pour les puces ADN à faible densité. Weigel et al. (Weigel et al., 2010) ont confirmé que la précision de la prédiction génomique ne diminue pas lors de l'utilisation de génotypes imputés si le panel de marqueurs de faible densité comprend plus de 3 000 SNPs uniformément répartis sur le génome. Burdick et al. (Burdick et al., 2006) ont ainsi proposé que les individus servant à la population d'entrainement des modèles utilisés pour la prédiction génomique soient génotypés sur une puce MD ou HD, tandis que les candidats à la sélection (le plus souvent apparentés aux individus de la population de référence) seraient génotypés sur une puce basse densité (BD) dont les SNPs sont régulièrement espacés. En utilisant la co-ségrégation des marqueurs de la puce

HD avec ceux de la puce BD au sein d'une famille, les génotypes de la puce HD sont ainsi imputés pour les candidats à la sélection (Figure 21). Ces génotypes imputés sont ensuite utilisés pour prédire les GEBV des candidats (Habier et al., 2009). Chez le porc, Zhang et al. (Zhang et al., 2018) ont montré que l'augmentation de la densité des marqueurs n'augmentait pas ou peu la précision des prédictions génomiques des caractères d'efficacité alimentaire chez les porcs Duroc. Ces résultats sont également rapportés par Song et al. (Song et al., 2019) qui indiquent que l'imputation d'une puce MD jusqu'à la séquence n'a pas augmenté la précision des prédictions des 3 populations de porcs Yorkshire étudiées indépendamment

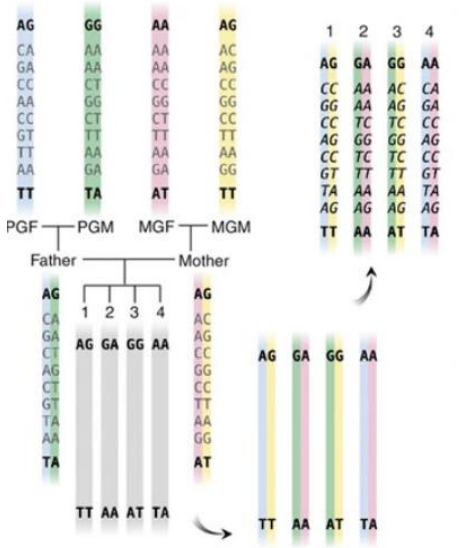

Figure 21 : Schéma explicatif de l'imputation dans le cadre de la prédiction génomique où les descendants sont génotypés sur une puce BD, et les parents et grand-parents sur une puce moyenne ou haute densité. PGF = Grand-père paternel ; PGM = Grand-mère paternelle ; MGF = Grand-père maternel ; MGM = Grand-mère maternelle ; Father = Père ; Mother = Mère. (Burdick et al., 2006)

3. Des régions génomiques identifiées pour l'efficacité alimentaire

a. Cartographie de QTL pour la RFI chez le porc

Parmi les différentes études d'association réalisées chez le porc, plusieurs articles ont été publiés sur l'étude de la CMJR et des caractères associés : GMQ, IC et CMJ. Dans les races porcines, relativement peu de QTL pour la CMJR ont été identifiés par rapport aux autres caractères de production. Les régions génomiques associées à ce caractère sont réparties sur l'ensemble du génome (Figure 22) et un bilan de l'ensemble des QTL identifiés pour l'efficacité alimentaire ont été reportés dans la table annexe 2. Parmi l'ensemble des QTL identifiés dans différentes études, le QTL localisé sur le SSC7 a été détecté dans deux études indépendantes (Do et al., 2014; Onteru et al., 2013). De plus, les régions génomiques mises en évidence présentent de petits effets qui semblent être spécifiques aux races et aux

populations de porcs analysées. La CMJR semble donc être un caractère hautement polygénique chez le porc.

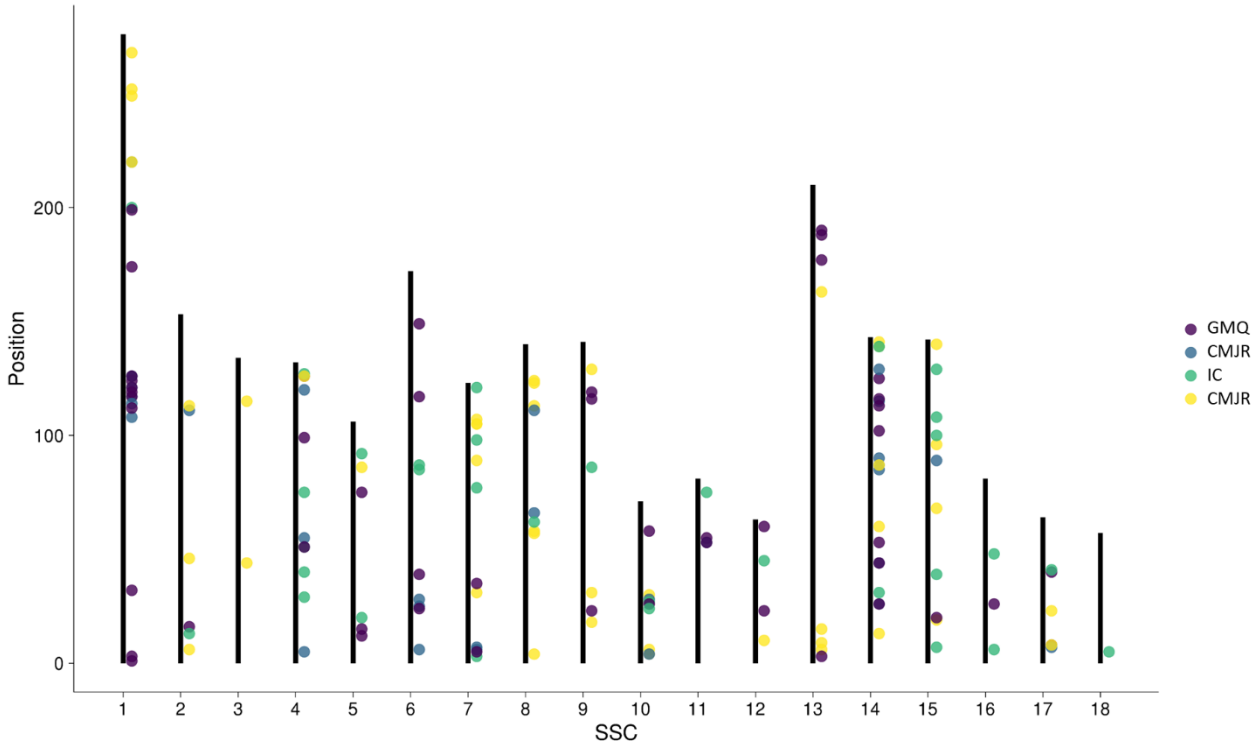

Figure 22 : Positionnement des QTL identifiés pour 4 caractères caractéristiques de l'efficacité alimentaire (CMJR, CMJ, IC et GMQ) et uniquement identifiés dans des études sur l'analyse de la CMJR chez des porcs en croissance. Seuls les 18 autosomes de porcs ont été représentés.

b. Cartographie chez les autres espèces de rente

La CMJR est un caractère qui a également été analysé pour d'autres espèces de rentes. Pour l'espèce bovine, 212 QTL ont été identifiés sur l'ensemble des autosomes (bovine QTLdb). Pour l'ensemble des études réalisées les résultats obtenus indiquent que ce caractère est contrôlé par de nombreux variants ayant des effets relativement faibles (Hu et al., 2019). En volaille, 14 régions QTL ont été identifiées pour le RFI dans la base de données chicken QTLdb (Hu et al., 2019). La région QTL identifiée sur le GGA27 a été retrouvée dans les deux principales études de l'efficacité alimentaire chez les poulets, avec notamment la détection d'un SNP significatif associé avec la CMJR (Yuan et al., 2015b).

III. Apport des données fonctionnelles à la cartographie de QTL

Alors que les caractères simples peuvent être expliqués que par quelques loci à effet fort, la ségrégation de nombreux loci à effet faible complique l'identification précise des variants causaux pour l'étude des caractères complexes (Bodmer and Bonilla, 2008). De ce fait l'apport des données fonctionnelles est rapidement apparu comme une source d'informations complémentaires pour contribuer à la caractérisation d'un phénotype particulier.

1. Les données transcriptomiques et leurs utilisations

a. Evolution des technologies d'analyse du transcriptome

Un transcriptome est la collection d'ARN exprimés dans une cellule ou un type de tissu spécifique, dans une condition donnée et à un moment donné. Le but ultime de la transcriptomique est de déterminer l'identité, l'abondance et la fonction de chaque transcrit dans chaque cellule ou type de tissu dans des conditions données, mais aussi de caractériser la façon dont les transcrits réagissent aux changements internes et/ou externes. Les premiers travaux ont porté sur le séquençage systématique, au sein d'un échantillon, d'un grand nombre de transcrits (EST : Expressed Sequence Tag). Cette approche nécessitait au préalable le clonage des différents cDNA de l'échantillon cible et la réalisation d'une collection de clones indépendants. Rapidement cette approche étaitremplacée au milieu des années 90 par des technologies basées sur l'hybridation d'une population d'ARN d'un tissu cible sur un support de référence. Le séquençage d'EST à grande échelle et le séquençage d'ADNc complet pour un grand nombre de tissus ont néanmoins fourni une énorme quantité de séquences de transcrits pour la fabrication de puces de référence à base d'ADNc et des puces à base d'oligo-sondes (Lockhart et al., 1996; Schena et al., 1995). Ces puces ont permis de quantifier simultanément l'abondance relative de milliers de transcrits (gènes) en les hybridant à des sondes fixées sur une surface solide, et en visualisant et quantifiant l'abondance de ces hybridations par des méthodes basées sur la fluorescence (Schena et al., 1995).

Les travaux pionniers de Su et al (Su et al., 2002, 2004) ont permis la première analyse complète du transcriptome codant pour les protéines des principaux organes de l'homme et de la souris, en réalisant le profil d'expression de 46 tissus humains et 45 tissus de souris d'origines diverses. Pour l'espèce porcine, des puces de différentes capacités ont été créés en fonction de l'intérêt des chercheurs et de la disponibilité des connaissances sur le transcriptome porcin, comme les puces (i) Agilent-037880/INRASus scrofa60K v1, enrichie en gènes s'exprimant dans le système immunitaire, les muscles et les tissus adipeux, (ii) Affymetrix Porcine Genome Array dont le choix des gènes est basé sur UniGene Build 28 (2004), (iii) AffymetrixPorGene-1_0-st-v1 et (iv) Affymetrix Porcine Snowball array (2010) (Schroyen and Tuggle, 2015).

Bien qu'utilisé pour de nombreuses études, cette technologie présente l'inconvénient de limiter l'analyse aux gènes présents sur la puce. En outre, cette technologie souffre de problèmes tels que le nombre fini de transcrits analysés, l'hybridation croisée et le bruit de fond (Draghici et al., 2006). L'essor des technologies de séquençage de nouvelle génération (NGS) a également révolutionné les études transcriptomiques à haut débit. Le séquençage de l'ARN a pour principal avantage de ne pas limiter l'analyse aux loci présents sur un support, et de ne pas nécessiter, comme prérequis, une connaissance du génome ou du transcriptome ciblé. De plus les résultats obtenus via cette technologie présentent un signal de bruit de fond très faible, voire inexistant, car les séquences d'ADN responsables de ce bruit de fond peuvent être cartographiées sans ambiguïté sur des régions uniques du génome. Enfin, le RNA-seq possède également comme avantage une reproductibilité plus élevée que les méthodes basées sur les puces à ADN (Wang et al., 2009b). Les applications de cette technologie ont été rapidement étendues depuis sa première utilisation pour l'acquisition des données d'expression génétique de la levure en 2008 (Nagalakshmi et al., 2008). Pour la recherche de gènes exprimés différentiellement (DEG) entre des conditions, les lectures de séquences d'ARN sont mises en correspondance avec le génome de référence, afin de comptabiliser le nombre de copie des gènes et exons de l'échantillon. Pour les espèces sans génome de référence de bonne qualité, un assemblage du transcriptome de novo pour obtenir un transcriptome de référence approximatif est réalisé, puis les lectures des séquences d'ARN sont alignées sur ce transcriptome de novo afin de procéder à la quantification. Cette technologie permet également une analyse très fine de la régulation du génome au sein d'une seule cellule et d'obtenir ainsi une mesure quantitative précise du transcriptome dans des cellules individuelles (Wu et al., 2014).

Les principaux objectifs de la transcriptomique sont les suivants : (i) cataloguer toutes les espèces de transcrits, que ce soit les ARNm, les ARN non codants et les petits ARN ; (ii) déterminer la structure transcriptomique des gènes en fonction de leurs sites d'initiation, de leurs extrémités 5' et 3', des modèles d'épissage et d'autres modifications post-transcriptionnelles ; (iii) et quantifier les niveaux d'expression changeants de chaque transcrit au cours du développement et/ou dans différentes conditions (Wang et al., 2009b). De plus, le transcriptome assemblé peut jouer un rôle important dans l'annotation du génome, la découverte de variants d'expression, le développement de marqueurs, la phylogénétique. Par conséquent, l'analyse transcriptomique des données via le RNA-seq permet de détecter de nouvelles régions transcrites ou la présence de formes alternatives d'épissage d'un gène. Cependant, l'identification de variants d'épissage à l'aide d'une approche de séquençage suppose qu'un nombre suffisant de lectures couvre les jonctions exon-exon (Marioni et al., 2008). Si peu à peu le RNA-seq devient la méthode privilégiée pour l'analyse d'un transcriptome, les puces ont

présenté pendant longtemps le principal avantage d'avoir un cout plus faible que les autres technologies et beaucoup de données ont été acquises à l'aide de cet outil (Malone and Oliver, 2011).

b. Les analyses transcriptome appliquées à la RFI chez le porc

i. Caractérisation physiologique

Chez le porc de nombreuses études transcriptomiques ont été publiées pour différents caractères d'intérêt chez cette espèce. Parmi ces travaux certains ont été réalisés afin de comprendre les mécanismes sous-jacent à la CMJR en utilisant les plus souvent des animaux issus des deux dispositifs de sélections divergentes présentés précédemment, et en appliquant des situations contrastées comme des restrictions alimentaires. Les tissus cibles analysés ont été le tissu adipeux, le foie, le muscle et le sang. En établissant le profil transcriptomique à l'aide de puces à ADN du tissu adipeux des deux lignées ISU de porcs sélectionnés de manière divergente pour la CMJR, Lkhagvadorj et al. (Lkhagvadorj et al., 2009) ont constaté que les gènes impliqués dans la voie du métabolisme des lipides étaient surreprésentés parmi les gènes différentiellement exprimés (DEG) (Figure 23). La majorité d'entre eux sont moins exprimés chez les porcs CMJR- que chez les porcs CMJR+, alors que les gènes différentiels impliqués dans le métabolisme des glucides et la réponse au stress étaient surexprimés chez les porcs CMJR-. Ils ont également constaté que le transcriptome d'un réseau de gènes associé à la leptine était différent entre les deux lignées. Les études de Vincent et al. (Vincent et al., 2015) et de Jing et al. (Jing et al., 2015) ont permis d'établir le profil du transcriptome du muscle longissimus dorsi (LD) chez des porcs issus respectivement des lignées divergentes INRAE et ISU au moyen de puces à ADN (Vincent et al. 2015) et de séquences d'ARN (Jing et al. 2015). Vincent et al. (Vincent et al., 2015) ont trouvé des différentiels d'expressions pour des gènes impliqués dans la synthèse des protéines et la glycolyse (surexpression dans la lignée CMJR-), ainsi que pour des gènes associés à l'énergie mitochondriale et au métabolisme oxydatif (sous-expression dans la lignée CMJR-). En revanche, en utilisant des porcs Yorkshire sélectionnés de manière divergente par rapport à la CMJR, Jing et al. (Jing et al., 2015) ont trouvé que les gènes impliqués dans la glycolyse avaient une expression plus faible pour les CMJR-, alors que les gènes impliqués dans la prolifération et la différenciation musculaires avaient une expression plus élevée pour les porcs CMJR-. Ils ont également découvert que les gènes différentiellement exprimés étaient en fait reliés directement ou indirectement par un ou deux gènes seulement. Dans le réseau (Figure 24), l'activité mitochondriale a été séparée en trois parties : la réaction de découplage, le contrôle respiratoire des mitochondries et la régulation de la transcription des mitochondries. Étonnamment, les études (Jing et al., 2015; Vincent et al., 2015) ne partagent aucun gènes différentiellement exprimés et proposent des différences métaboliques entre lignées complètement opposées. En 2017, Gondret et al. (Gondret et al., 2017) ont analysé les profils transcriptomiques de quatre tissus impliqués dans la production et l'utilisation de l'énergie chez les

porcs, correspondant au foie et à deux tissus adipeux en complément du muscle de la longe des 48 porcs des deux lignées divergentes sélectionnées par rapport à leur CMJR (24 CMJR+ et 24 CMJR-). Dans cette étude, les auteurs se sont concentrés sur les voies moléculaires communes et spécifiques des différents tissus et les résultats mis en évidence sont que la réponse immunitaire, le métabolisme des protéines et la réponse au stress oxydatif seraient les principales voies moléculaires associées à une différence génétique pour la CMJR. Ils ont également émis l'hypothèse que les fonctions non productives dans les tissus métaboliques seraient probablement des processus importants pour l'efficacité alimentaire. Sur ce même dispositif, l'analyse du transcriptome sanguin entier était une approche intéressante pour mieux comprendre les mécanismes moléculaires qui sous-tendent les différences de CMJR, pour déterminer les relations possibles entre les caractéristiques du sang et les caractères de production, et plus généralement pour identifier des gènes facilement accessibles afin de suivre des changements physiologiques en réponse à des facteurs particuliers. Jégou et al. (Jégou et al., 2016) ont ainsi étudié le transcriptome sanguin des 48 porcs sélectionnés pour leurs valeurs de CMJR, et les auteurs n'ont pas obtenu de différence significative dans le nombre de lymphocytes, de monocytes et de neutrophiles entre les lignées divergentes. Ces observations diffèrent de l'étude de Mpetile et al. (Mpetile et al., 2015) faisant état d'un nombre plus faible de globule blanc, en particulier de lymphocytes et de monocytes, chez les porcs Yorkshire sélectionnés sur la CMJR. En ce qui concerne les globules rouges, Jégou et al. (Jégou et al., 2016) ont noté que leur nombre total était plus élevé chez les porcs CMJR- que chez les porcs CMJR+, alors que ce nombre ne différait pas entre les lignées dans l'étude de Mpetile et al. Il est important de noter que d'autres paramètres liés aux globules rouges ont été affectés de manière similaire par la sélection sur la CMJR dans les deux études, avec une concentration plasmatique d'hémoglobine plus élevée et un pourcentage d'hématocrite plus élevé chez les porcs CMJR- que chez les porcs CMJR+. Jégou et al. indiquent que la divergence des résultats pourrait refléter des différences au niveau de l'intensité de la sélection, de l'âge des porcs pris en compte pour l'étude des cellules sanguines ou encore les différents stimuli environnementaux pouvant avoir eu lieu durant l'élevage des porcs.

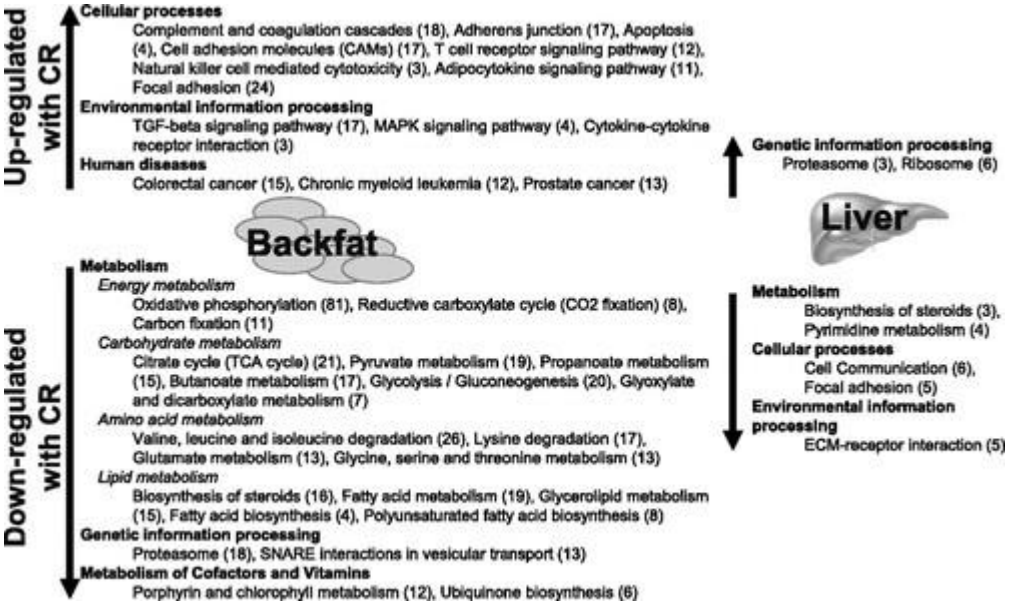

Figure 23 : Résultat général de l'étude d'annotation fonctionnelle de la réponse du foie et du tissu adipeux au jeûne, en identifiant les voies biologiques KEGG surreprésentées (P<0,05) en gènes différentiellement exprimés par Lkhagvadorj (Lkhagvadorj et al., 2009)

D'autre part les études transcriptomiques peuvent être un outil intéressant pour élaborer des listes de biomarqueurs candidats pour un phénotype d'intérêt. Dans cette optique, Liu et al. (Liu et al., 2016) ont réalisé la prédiction de biomarqueurs pour la CMJR à partir de données transcriptomiques du sang post-sevrage issus de porcs Yorkshire sélectionnés pour leurs valeurs de CMJR. Les DEG dont les niveaux d'expression étaient associés au phénotype CMJR, ainsi que les gènes dont les profils d'expression étaient similaires aux gènes des modules de co-expression associés à la CMJR ont été considérés comme de bons candidats pour des biomarqueurs de la CMJR. D'après leurs résultats, cinq biomarqueurs candidats ont été sélectionnés dont LRP6 et PDL1. De plus, Messad et al. (Messad et al., 2019) ont également identifié des gènes exprimés dans le muscle qui pourraient être des prédicteurs fiables de l'efficacité alimentaire, à partir des données transcriptomiques de longe issus d'animaux du dispositif de sélection divergente INRAE. Pour chaque porc comportant des données d'expression, la valeur d'élevage pour la CMJR a été estimée et l'indice de consommation a été calculé pendant les périodes de test des animaux. La procédure de régression progressive a mis en évidence 10 gènes (FKBP5, MUM1, AKAP12, FYN, TMED3, PHKB, TGF, SOCS6, ILR4 et FRAS1) dans une combinaison linéaire prédisant l'IC. En outre, FKBP5 et les niveaux d'expression de cinq autres gènes (IGF2, SERINC3, CSRNP3, EZR et RPL16) ont contribué de manière significative à la valeur d'élevage pour la CMJR.

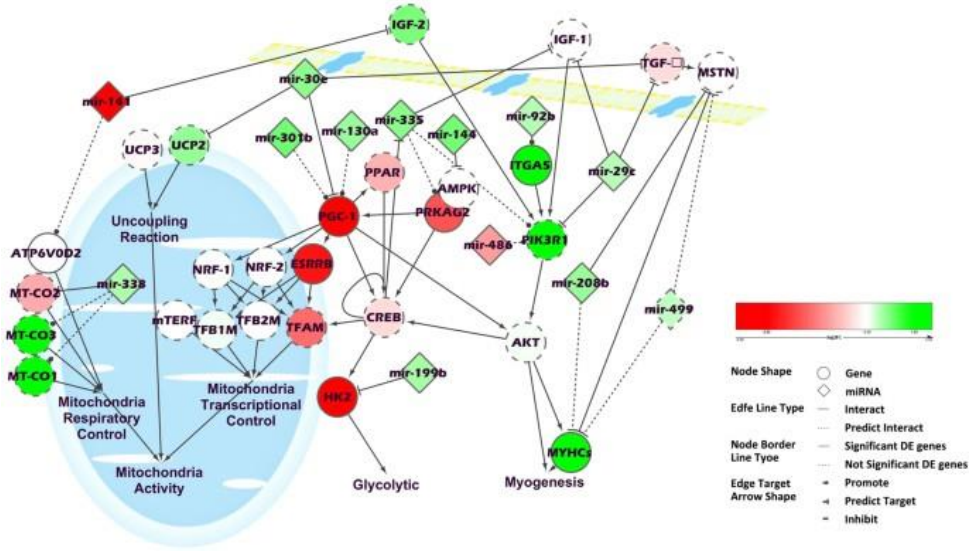

Figure 24 : Le réseau clé de gènes et d'ARNmi s'est avéré être exprimé différemment dans le muscle squelettique des porcs CMJR- par rapport à celui des porcs CMJR+. Le réseau a été réalisé à l'aide de Cytoscape. (Jing et al., 2015)

ii. Utilisation comme phénotype (eQTL)

La génomique génétique, en tant que nouvelle approche qui combine les données d'expression des gènes et les génotypes marqueurs dans une population en ségrégation, offre de grandes perspectives pour apporter une contribution majeure à la dissection de caractères complexes (Jansen, 2003; Jansen and Nap, 2001). La génomique génétique vise à améliorer les connaissances sur les voies moléculaires sous-jacentes aux régions QTL identifiées par des études GWAS (aussi nommée eQTL pour expressionQTL). Les analyses eQTL (expression QTL) suggèrent donc que les variants génétiques contribueraient au caractère d'intérêt *via* la perturbation de l'expression des gènes (Civelek and Lusis, 2014). L'eQTL détectée peut représenter un locus proche du gène contrôlé (cis-acting) ou un ou plusieurs loci non liés au gène contrôlé (trans-acting) (Jansen and Nap, 2001). Le eQTL pour les gènes présentant une forte corrélation avec le phénotype peut fournir les informations nécessaires à l'identification des gènes qui contrôlent les phénotypes quantitatifs analysés en GWAS. Ces cis-eQTL résultant de la corrélation des profils d'expression avec les mesures phénotypiques représentent des gènes candidats pour la régulation génétique qui sous-tend la variation des caractères physiologiques (Doss et al., 2005; Schadt et al., 2005). Chez l'homme, plus de 40 % des variants associés à une maladie ont un effet *cis* sur l'expression des gènes (Nica et al., 2010). En complément, les *trans*-eQTL peuvent fournir des informations sur les voies et mécanismes moléculaires aboutissant à une maladie (Fehrmann et al., 2011). Ponsuksili et al. (Ponsuksili et al., 2010, 2014, 2015) ont réalisé plusieurs approches eQTL pour différents caractères mesurés chez les porcs, comme notamment la première analyse complète de *cis*-eQTL associés aux caractéristiques de qualité de la viande chez le porc permettant d'établir une liste de gènes candidats pour ces caractères (Ponsuksili et al., 2008). Ensuite

une seconde approche basée cette fois sur l'analyse du muscle de la longe et des caractères de qualité de viande, a donné une meilleure caractérisation des relations complexes présentent entre les différents tissus étudiés. Ainsi, les processus biologiques qui sont détectés dans le muscle squelettique et la qualité de la viande sont déterminées d'une part par les réserves d'énergie et leurs utilisations dans le muscle, et d'autre part par la structure musculaire elle-même et la signalisation du calcium (Ponsuksili et al., 2010). Une première analyse de cis et *trans*-eQTL, dans une population de Danbred Durocs (N=11) et Danbred Landrace (N=27), a été réalisée par Carmelo et al. (Carmelo and Kadarmideen, 2020). Les auteurs ont identifié 15 eQTL avec un FDR < 0,01, affectant plusieurs gènes trouvés dans des études précédentes de races porcines commerciales. Il s'agit par exemple des gènes IFI6, PRPF39, TMEM222, CSRNP1, PARK7 et MFF. De plus, l'analyse de l'enrichissement des *trans*-eQTL a révélé un enrichissement élevé pour les gènes et les ontologies de gènes associés au contexte génomique et à la régulation de l'expression comme par exemple les facteurs de transcription, la liaison à l'ADN, l'activité du facteur de transcription de liaison à l'ADN, la régulation positive de l'expression, ou encore la régulation négative de l'expression.

2. L'annotation fonctionnelle pour la cartographie de régions génomiques

a. De l'évolution des connaissances vers leurs mises en commun

Les développements méthodologiques de biologie moléculaire visant à mieux comprendre le fonctionnement des génomes sont à l'origine de la mise en place d'initiatives internationales dans ce domaine. Ainsi, dans la lignée des projets de séquençage (International Human Genome Sequencing Consortium, 2001) et d'haplotypage du génome humain (the international HapMap project (Gibbs et al., 2003)), le projet ENCODE (ENCyclopedia Of DNA Elements, https://www.encodeproject.org/) a vu le jour en 2003. L'objectif était alors d'identifier les éléments fonctionnels dans la séquence du génome humain tout en partageant et homogénéisant les protocoles expérimentaux, les jeux de données et les analyses bioinformatiques afin de garantir une reproductibilité et une utilité à la communauté scientifique internationale (Dunham et al., 2012).

Le projet ENCODE initié au début des années 2000 a été organisé en 4 phases. La première étape, ENCODE I, consistait en une étude pilote visant à tester et comparer différentes méthodes sur une fraction de 1% du génome humain dont la moitié a été sélectionnée pour contenir des éléments régulateurs connus. L'identification à grande échelle d'une variété d'éléments fonctionnels, comme des gènes, des promoteurs, des amplificateurs, répresseurs/silencieux, des exons, des sites de

terminaison de la réplication, des sites de liaisons des facteurs de transcription et des séquences conservées par plusieurs espèces (Figure 25) a ainsi pu être possible (Birney et al., 2007).

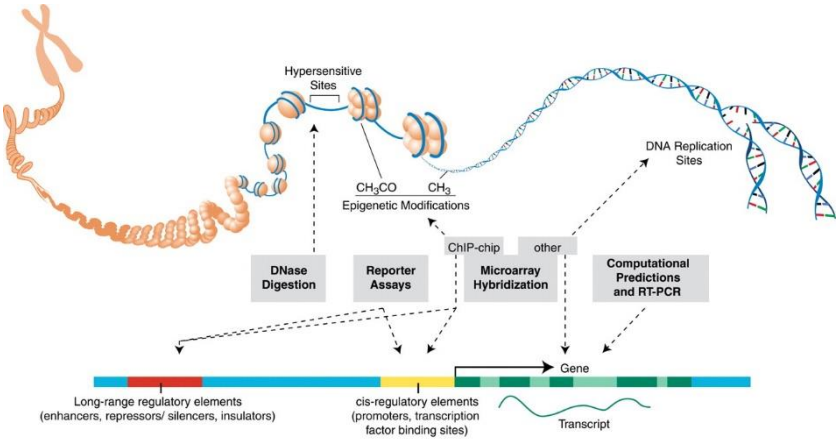

Figure 25 : Les éléments génomiques fonctionnels identifiés par la phase I d'ENCODE. Les méthodes indiquées sont utilisées pour identifier différents types d'éléments fonctionnels dans le génome humain. (Consortium, 2004)

La seconde phase du projet, ENCODE II, a consisté à générer et analyser des données sur le génome complet afin d'aboutir à une caractérisation très résolutive des éléments fonctionnels du génome humain. Cette étape a été rendue possible grâce aux technologies basées sur le séquençage pour l'acquisition de divers données -omiques (Dunham et al., 2012) (Figure 26). En effet, 1 640 jeux de données incluant 147 types cellulaires différents ont été produits et intégrés avec d'autres ressources telles que des régions génomiques issues des GWAS montrant, entre autres, que les variants associés avec des maladies se retrouvent en majorité dans des éléments fonctionnels non-codants (Dunham et al., 2012).

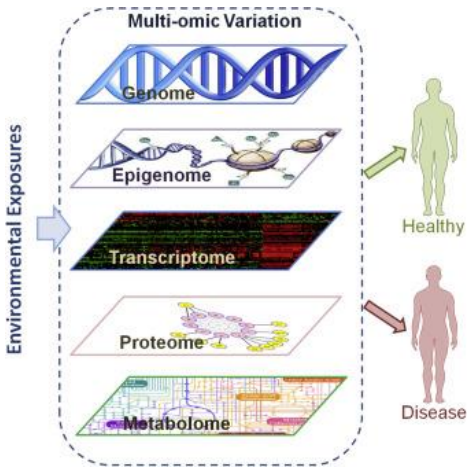

Figure 26 : Acquisition de données omics à différentes échelles pour obtenir des données sur l'ADN, l'ARN, les protéines et les métabolites, afin de mieux comprendre la biologie de caractères complexes. (Sun and Hu, 2016)

La troisième phase du projet, ENCODE III, a été dans la continuité d'ENCODE II mais impliquant un aspect développemental et évolutif sur la connaissance de la régulation des génomes *via* la mise en place d'un équivalent au projet humain mais sur le génome murin (Figure 27) (Snyder et al., 2020). Ici, 5 992 nouveaux jeux de données ont été produits parmi lesquels des données acquises chez la souris *in utero* afin d'évaluer la régulation du génome au cours du développement embryonnaire (Moore et al., 2020). L'intégration des données disponibles a permis de générer un répertoire de 926 535 (homme) et 33 815 (souris) éléments de régulation couvrant respectivement 7,9 et 3,4% des génomes humains et murins (Moore et al., 2020).

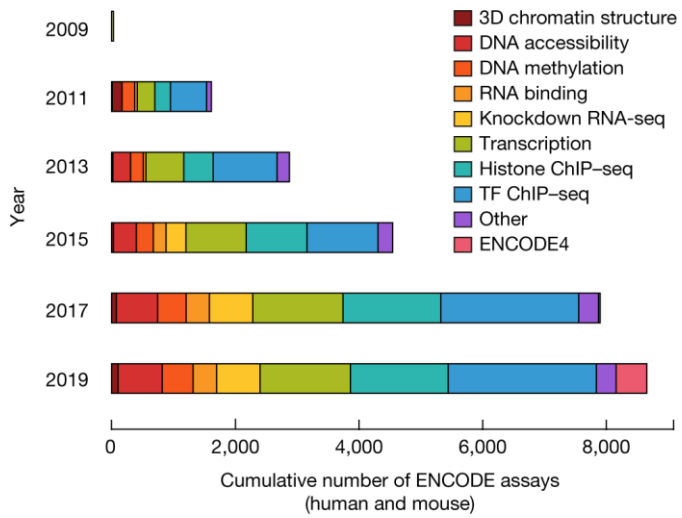

Figure 27 : Accumulation de nouveaux types d'analyses au cours des trois phases d'ENCODE. (Snyder et al., 2020)

Aujourd'hui, la quatrième phase du projet, ENCODE IV, a débuté avec comme objectif d'améliorer les connaissances dans les régions, types cellulaires ou tissus qui ont fait défaut dans les 3 premières phases. C'est pourquoi de nouvelles expérimentations à partir d'échantillons rares (tissu fœtal humain et organes reproducteurs humains) autour de la transcriptomique sur cellule unique ou de la caractérisation des isoformes pleines longueurs des ARNm sont en cours (Snyder et al., 2020). En terme d'analyses, un focus sera fait sur la spécificité allélique que ce soit pour l'expression des ARNm, la méthylation ou la fixation de facteurs de transcription (Yang et al., 2019). Bien qu'initié il y a 20 ans, le consortium ENCODE a encore une grande place dans le domaine scientifique, comme en atteste les courbes de publications issus du consortium lui-même ou encore le nombre de publications utilisant les donnant provenant du projet ENCODE (Figure 28).

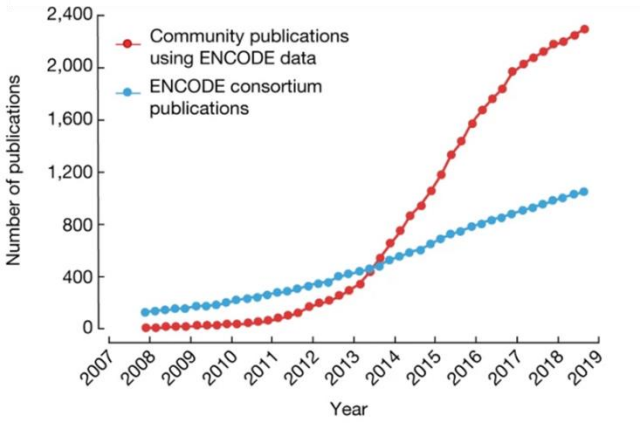

Figure 28 : L'Institut national de recherche sur le génome humain (NHGRI) a identifié une liste de publications qui ont utilisé les données ENCODE. Les publications communautaires (« Community publications ») sont identifiées par des recherches automatisées pour la citation des numéros d'adhésion ENCODE, des documents phares ENCODE ou des ressources telles que HaploReg et RegulomeDB et manuellement pour déterminer si les données ENCODE ont été effectivement utilisées. Les documents du consortium (« ENCODE consortium ») sont identifiés par des recherches automatisées dans PubMed pour les publications qui ont été soutenues au moins en partie par des prix ENCODE. (Moore et al., 2020)

Les résultats de ces études ont permis l'établissement de diverses cartes d'éléments régulateurs (promoteurs, amplificateurs, répresseurs, insulateurs) dans les génomes humains et murins mais aussi d'établir une dynamique de leurs mises en place au cours du développement. L'accès à tous ces éléments fonctionnels permet ainsi d'améliorer les connaissances sur la régulation du génome mais aussi d'établir des liens de causalité de ces éléments régulateurs avec des maladies *via* l'annotation fonctionnelle des variants (Pazin, 2015; Sun and Hu, 2016). Ainsi, l'intégration de toutes ces informations omiques représente aujourd'hui un enjeu majeur en biologie pour mieux comprendre et affiner l'architecture moléculaire des maladies complexes telles que les cancers (Zhang et al., 2020) ou polygéniques comme le diabète (Sridhar, 2018). Depuis, d'autres projets de grande ampleur dont le but est de fédérer la communauté internationale autour d'une thématique ont vu le jour, comme par exemple le programme de l'Atlas du génome du cancer (TCGA) (Tomczak et al., 2015), le Consortium international de l'épigénome humain (IHEC) (Stunnenberg et al., 2016), le Genotype-Tissue Project (GTEx) (Lonsdale et al., 2013) pour l'homme, ou encore le Functional Annotation of Animal Genomes (FAANG) (Andersson et al., 2015) chez les animaux de rente.

b. Le consortium FAANG pour les animaux de rentes

Le consortium international du projet FAANG (Functional Annotation of ANimal Genomes) a été lancé en 2015 (http://www.faang.org) afin de rassembler les scientifiques travaillant dans les domaines de la génétique et génomique chez les animaux d'élevage. La recherche sur les animaux domestiques a des répercussions scientifiques et socio-économiques importantes, notamment en contribuant à la recherche médicale et en soutenant les améliorations dans le secteur de l'élevage. La richesse de la

diversité génétique et phénotypique des animaux d'élevage représente un élément clé des impacts de ce projet afin d'élucider l'architecture moléculaire des caractères agronomiques d'intérêt (Andersson et al., 2015). Les mutations dans les gènes *IGF2* (Van Laere et al., 2003) et *DLK1* (Freking et al., 2002), identifiées respectivement dans les espèces porcine et ovine, sont localisées dans des régions génomiques non-codantes et affectent, *via* des mécanismes épigénétiques, la musculature des animaux. Ces exemples illustrent donc l'intérêt de déterminer des cartes d'éléments fonctionnels pour les génomes animaux.

Le plan de route établi dans le cadre du projet FAANG est de générer des jeux de données pangénomiques (Figure 29), incluant expression des ARNs, méthylation de l'ADN, modification et accessibilité de la chromatine et interactions chromosomiques, et ce sur 8 espèces majeures (*Bos taurus, Bos indicus, Bubalus bubalis, Capra Hircus, Equus caballus, Gallus gallus, Ovis aries et Sus scrofa*) (Giuffra and Tuggle, 2019).

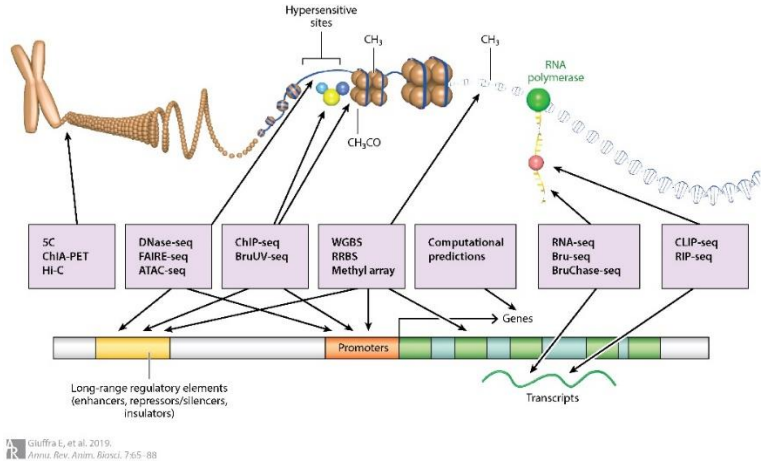

Figure 29 : Annotation fonctionnelle des génomes par l'analyse d'ensembles de données obtenues par différentes approches de séquençage a permis de construire des cartes complètes d'annotation du génome, comprenant des éléments qui agissent au niveau des protéines et de l'ARN et des éléments de régulation. (Giuffra and Tuggle, 2019)

En octobre 2018, environ 400 scientifiques internationaux étaient mentionnés comme contributeurs à ce projet et 9 358 expérimentations avaient été réalisées répondant aux prérequis définis puis mises à disposition sur le portail en ligne (https://data.faang.org/home). Parmi les premiers résultats issus de ces travaux, une comparaison multi-espèces (bovine, caprine, porcine et aviaire) des éléments fonctionnels identifiés à partir d'expérimentations multi-omiques de transcription, de conformation et d'accessibilité de la chromatine dans 3 tissus représente une avancée majeure pour améliorer les connaissances sur les génomes animaux (Foissac et al., 2019). Les analyses comparatives ont montré *(i)* qu'un ensemble de régions régulatrices sont conservées chez les différentes espèces et *(ii)* que les limites des domaines d'association topologique (TADs) conservées entre espèces ont des

propriétés d'isolation plus fortes que celles spécifiques à l'espèce (Foissac et al., 2019). Tout comme les projets ENCODE, les cartes d'éléments fonctionnels du projet FAANG ont pour objectif de générer une ressource de données qui sera utilisée par la communauté scientifique internationale, à long terme et à des fins multiples (Figure 29). En outre, ces nouvelles connaissances ouvrent la voie aux approches de biologie des systèmes et de génétique systémique chez les animaux d'élevage afin de comprendre quelles sont les grandes fonctions biologiques impliquées dans la variabilité des caractères agronomiques (Suravajhala et al., 2016). C'est dans ce cadre que les projets européens Gene-Switch (https://www.gene-switch.eu) et BovReg (https://www.bovreg.eu/) basés en partie sur l'exploitation des données FAANG ont été développés.

c. L'ontologie génétique et ses outils

La multitude d'informations acquises sur les gènes et les protéines depuis les prémisses du séquençage a démontré qu'une grande partie des gènes et de leurs fonctions était conservée chez les eucaryotes. En revanche, une nomenclature uniforme de l'annotation des gènes et des protéines faisait alors défaut rendant ainsi difficile les comparaisons entre espèces. C'est dans cette optique que le consortium Gene Ontology (GO) (http://geneontology.org/) a été créé au début des années 2000 afin de produire une nomenclature homogène, partagée et dynamique pour les gènes et protéines de divers organismes (Ashburner et al., 2000). Cette ontologie génétique représente une classification structurée (en termes et concepts) et contrôlée. Elle permet d'unifier la multiplicité des termes employés pour décrire un concept ce qui améliore l'interopérabilité entre toutes les bases de données. Au 1er janvier 2021, le portail de ressources regroupe 44091 GO termes, 934369 annotations pour 1561738 protéines issus de 4743 espèces différentes (http://geneontology.org/).

Chaque gène et son produit sont définis par 3 catégories de GO termes décrivant des entités biologiques sous trois aspects (processus biologique, fonction moléculaire et compartiment cellulaire). Ils peuvent *(i)* être adressés à un ou plusieurs compartiment cellulaires, *(ii)* participer à un ou plusieurs processus biologiques et *(iii)* y remplir une ou plusieurs fonctions moléculaires (Ashburner et al., 2000). Une annotation GO standard est faite en associant un produit génique à un terme GO représentant au mieux son rôle. Les niveaux d'annotation, directement issus de l'exploitation de la littérature scientifique, sont précisés par des codes également répartis en catégories selon qu'ils proviennent de résultats issus d'expérimentations biologiques, de la biologie numérique ou computationnelle (Poux and Gaudet, 2017). Actuellement, la base de données GO tire parti de données expérimentales provenant de plus de 140000 articles scientifiques, ce qui correspond à environ 600000 annotations basées sur des résultats biologiques (http://geneontology.org/). La base de données GO est structurée à l'aide d'une ontologie formelle qui ont des relations spécifiques les unes avec les autres (figure 30).

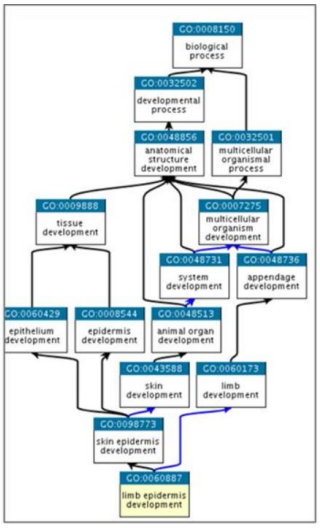

Figure 30 : Représentation d'un ensemble de termes GO décrite via un graphe, où chaque terme GO est un nœud, et les relations entre les termes sont des arêtes entre les nœuds. Par exemple, le terme de processus biologique "processus biosynthétique de l'hexose" a deux parents, "processus métabolique de l'hexose" et "processus biosynthétique des monosaccharides" (http://geneontology.org/docs/ontology-documentation/).

La classification GO est en développement continu depuis 20 ans, sans aucun signe de ralentissement. L'ontologie et les annotations continuent d'être régulièrement mises à jour, en réponse aux nouvelles découvertes expérimentales concernant la fonction des gènes, et à l'accumulation de connaissances sur la façon dont les gènes fonctionnent ensemble dans des systèmes biologiques plus vastes. Le consortium GO redouble d'efforts pour revoir les annotations, en particulier celles qui sont plus anciennes et qui pourraient être remplacées par des plus récentes (The Gene Ontology Consortium, 2019). Cette ressource représente donc un outil puissant pour l'analyse des données et les prévisions fonctionnelles de gènes à mesure que les ontologies et les annotations ont évolué (Bard and Rhee, 2004). Cependant, son utilisation nécessite de maitriser la structure de l'ontologie afin de ne pas en tirer des interprétations et conclusions erronées (Yon Rhee et al., 2008). Bien que les niveaux d'annotations provenant de données computationnelles améliorent la significativité de la classification, ils contiennent probablement plus de faux-positifs par rapport aux annotations issues directement de résultats biologiques expérimentaux (Yon Rhee et al., 2008).

Aujourd'hui, l'ontologie génétique est l'une des ressources informatiques les plus utilisées dans le domaine de la biologie des systèmes afin d'évaluer la fonction d'un gène non plus de façon ciblée mais intégrée dans un ou plusieurs processus biologiques (The Gene Ontology Consortium, 2019). Ces analyses reposent sur des méthodes statistiques dites d'enrichissement de gènes ou de voies de signalisation (Reimand et al., 2019). Parmi les différentes méthodes développées et intégrées dans des logiciels, l'outil GSEA (Gene Set Enrichment Analysis) a d'ores et déjà été largement utilisé pour des analyses transcriptomiques (Mootha et al., 2003; Subramanian et al., 2005). Les gènes sont ainsi classés

et ordonnés en fonction d'une valeur comme par exemple, le différentiel d'expression entre 2 conditions expérimentales (Figure 31a). L'hypothèse nulle est définie par le fait que les gènes appartenant à un même GO terme ne présentent aucune corrélation avec la liste ordonnée issue des analyses d'expression. L'autre hypothèse est qu'une annotation d'un GO terme est enrichie en gènes identifiés et rangés selon le différentiel d'expression. Ainsi, la méthode GSEA comporte trois éléments clés : *(i)* le calcul d'un score d'enrichissement (ES) qui correspond à une statistique pondérée de type Kolmogorov-Smirnov (Nonparametric Statistical Methods, 3rd Edition) (Figure 31b), *(ii)* l'estimation de la significativité de ES *via* l'utilisation d'un test de permutation empirique basé sur les phénotypes et *(iii)* l'ajustement pour les tests multiples *via* la normalisation de ES (NES) pour chaque ensemble de gènes et le calcul du FDR (False Discovery Rate) correspondant à chaque NES (Subramanian et al., 2005).

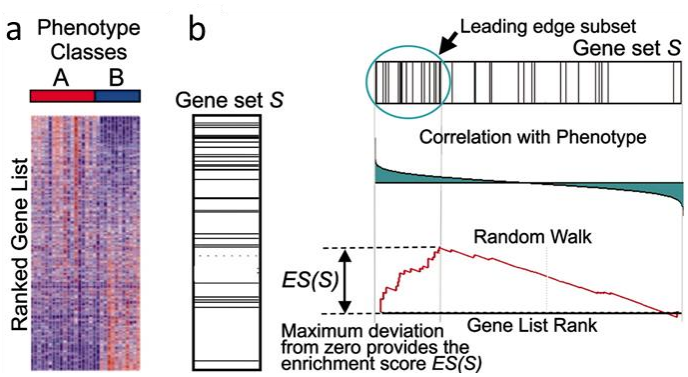

Figure 31 : Illustration de la méthode GSEA. (a) Un ensemble de données d'expression triées par rapport au différentiel d'expression obtenu entre les deux phénotypes étudiés et la carte thermique correspondante à ces valeurs, c'est-à-dire l'emplacement des gènes d'un ensemble S dans la liste triée. (b) Graphique de la somme courante pour S dans l'ensemble de données, y compris l'emplacement du score d'enrichissement maximal (ES). (Subramanian et al., 2005)

La méthode GSEA tirant partie de la base de données GO a principalement été utilisée pour l'analyse de données transcriptomiques chez plusieurs espèces et pour différents mécanismes biologiques démontrant ainsi la puissance de cet outil. A titre d'exemple, cette approche a été employée chez *Physcomitrella patens*, espèce de mousse utilisée comme organisme modèle pour les études sur l'évolution des plantes, afin d'évaluer la réponse à un stress abiotique causé par exemple par un changement d'environnement (Subramanian et al., 2005). Les analyses d'enrichissement s'avèrent efficaces là où *(i)* les corrections pour des tests multiples diminuent drastiquement le nombre de gènes statistiquement significatifs, *(ii)* la liste des gènes significatifs est très dense rendant difficile l'identification d'un dénominateur commun, *(iii)* la comparaison entre 2 listes de gènes statistiquement significatifs présente peu de chevauchement. Alors que des analyses focalisées sur des gènes candidats ne montraient que très peu de similitudes entre 2 études indépendantes sur des patients affectés

d'adénocarcinomes du poumon, les analyses GSEA ont révélé plusieurs voies biologiques en commun *via* l'enrichissement, entre autres, pour un des ensembles de gènes impliqués dans la signalisation de l'insuline, la synthèse de tRNA (Subramanian et al., 2005). Ainsi, l'essor des technologies omiques (principalement transcriptomique et protéomique) a été couplée à des développements méthodologiques (informatique et statistique), ce qui permet aujourd'hui d'aborder les questions de recherche autour de l'architecture moléculaire des caractères complexes à l'échelle du système ou de la fonction biologique.

d. Approches novatrices pour la caractérisation des régions génomiques en ségrégation à l'aide de données fonctionnelles

Jusqu'à présent, les études GWAS ont permis d'identifier des milliers de régions génomiques contribuant à la variabilité de caractères complexes mesurés sur divers organismes vivants. Bien que très répandues et très puissantes, ces approches présentent cependant des limites qui sont un frein à l'identification des relations génotype-phénotype (Tam et al., 2019). En effet, les études d'association pangénomiques *(i)* sont pénalisées par l'utilisation de méthodes de corrections statistiques pour assurer la robustesse et significativité des résultats, *(ii)* expliquent une faible proportion de la variance phénotypique de la plupart des caractères complexes, *(iii)* ne permettent ni la détermination précise des gènes ou variants causaux ni l'identification de la totalité des polymorphismes impliqués dans le déterminisme génétique des phénotypes et *(iv)* sont sensibles à la structure des populations (Tam et al., 2019). Ainsi un des enjeux de l'ère post-GWAS repose sur le développement de nouvelles approches permettant d'exploiter au mieux les résultats issus de ces études afin d'appréhender plus finement les mécanismes biologiques mis en jeu dans l'élaboration des caractères complexes. Parmi les diverses pistes d'amélioration proposées, certaines s'appuient sur le développement de modèles statistiques alternatifs (interactions G*E, épistasie, dominance/récessivité) (Cantor et al., 2010), d'autres consistent à augmenter les jeux de données *via* le recours à des phénotypes nouveaux ou des méta-analyses (Huang et al., 2015). Depuis quelques années, une autre stratégie basée sur l'intégration de données multi-omiques, appelée génétique systémique, est en plein essor (Civelek and Lusis, 2014).

L'avènement des technologies haut-débit permet d'étudier quantitativement des centaines ou des milliers de molécules biologiques, des variants d'ADN jusqu'aux marques épigénétiques, en passant par les niveaux de transcrits, de protéines et de métabolites. Ainsi, les approches de génétique systémique visent à tirer parti de ces phénotypes moléculaires intermédiaires pour établir un lien entre génome et phénome. Aujourd'hui, l'intégration des données génétiques avec différents types de données omiques permet de comprendre l'architecture moléculaire des caractères à l'échelle du système. Cette stratégie de génétique systémique représente donc un réel défi pour l'exploitation optimale des études d'associations pangénomiques puisqu'elle présente l'avantage d'évaluer les

interactions moléculaires dans un contexte qui est le plus pertinent pour les caractères complexes, à savoir des perturbations génétiques multiples (Civelek and Lusis, 2014) (Figure 32).

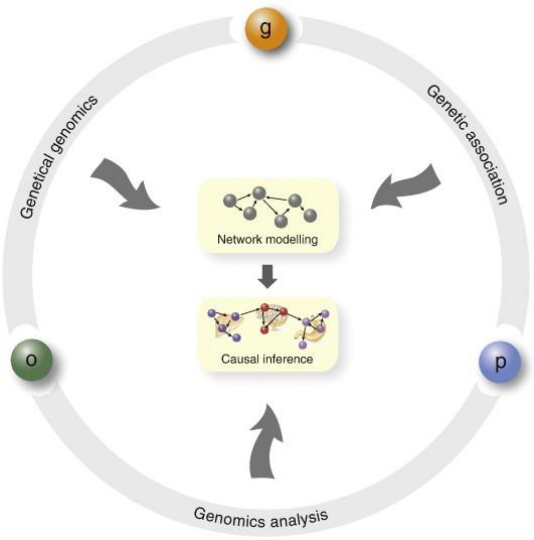

Figure 32 : La génétique des systèmes combine la génomique génétique, l'association génétique et les analyses génomiques pour construire l'inférence causale de génotype à phénotype. L'approche intègre les informations sur le génotype (g), de la ~omique (o) et les données sur le phénotype (p) et peut être utilisée pour construire un réseau et déduire la causalité des différents variants. (van der Sijde et al., 2014)

Jusqu'à présent, les approches de génétique systémique ont principalement été développées chez l'homme (van der Sijde et al., 2014) ou des organismes modèles comme les plantes (Feltus, 2014). En effet, chez ces espèces, l'accès à des bases de données fonctionnelles issues de projets internationaux comme ENCODE (https://www.encodeproject.org/), GTEx (https://gtexportal.org/home/) ou 10KP (https://db.cngb.org/10kp/) contribue à la mise en place d'une telle stratégie basée sur les connaissances des fonctions biologiques des gènes. Les analyses d'enrichissement de type GSEA, largement utilisées sur les données transcriptomiques, ont été depuis détournées pour l'exploitation de données génétiques issues de GWAS (Holden et al., 2008) et implémentées dans différents outils comme MAGENTA (http://www.broadinstitute.org/mpg/magenta) ou i-GSEA4GWAS V2 (http://gsea4gwas-v2.psych.ac.cn/). Ces méthodes permettent de déterminer si une liste définie de gènes, ici localisés dans les régions génomiques identifiées par les études d'associations pangénomiques, est statistiquement enrichie en groupe de gènes appartenant à un même terme GO (Zhang et al., 2015). L'intérêt de ces approches a été démontré, entre autres, pour *(i)* la maladie de Crohn en identifiant la voie IL12/IL23 associée avec la maladie (Torkamani et al., 2008; Wang et al., 2009a) ou *(ii)* plus récemment pour la fibrillation auriculaire mettant en avant l'importance de la voie de signalisation mTOR dans le développement de la pathologie (Ebana et al., 2019). Une façon complémentaire pour

prioriser et sélectionner des gènes candidats de données GWAS repose sur la connaissance de réseaux de gènes afin d'organiser et structurer les données en unités et interactions biologiquement significatives (Civelek and Lusis, 2014; Taşan et al., 2015). La modélisation de réseaux a été appliquée à un large éventail de questions biologiques et a contribué à la découverte de plusieurs gènes et biomarqueurs de maladies. A titre d'exemple, la comparaison de réseaux de co-expression de gènes entre des individus obèses et contrôles combinée aux résultats de GWAS a permis de mettre en évidence le gène *NEGR1* comme central impliqué (Walley et al., 2012). De plus en plus, les données d'annotations fonctionnelles permettent d'orienter et sélectionner les variants d'intérêt au sein d'une région génomique associée à un caractère s'affranchissant ainsi du fort déséquilibre de liaison qui existe entre tous les marqueurs de l'intervalle (Schaub et al., 2012). L'intégration des données issues des projets ENCODE (https://www.encodeproject.org/) et RoadMap Epigenomics (http://www.roadmapepigenomics.org/) a démontré l'importance significative de quelques éléments régulateurs non-codants pour certaines pathologies cardiaques (Villar et al., 2020). Dans une étude de grande ampleur visant à annoter les variants génétiques non-codants associés aux maladies cardiovasculaires à partir de cartes de conformation de la chromatine et de données de méthylation et acétylation des histones, Montefiori et al. ont montré qu'environ 90% des interactions entre SNPs associés et gènes cibles probables n'impliquaient pas le gène le plus proche (Montefiori et al., 2018). Ces résultats soulignent l'importance de la régulation à longue distance pour l'interprétation fonctionnelle des variants génétiques non-codants.

Une telle approche de génétique systémique chez les animaux d'élevage peut aujourd'hui être envisagée grâce à l'acquisition de jeux de données multi-omiques combinée aux efforts de la communauté internationale sur l'amélioration des connaissances des génomes animaux, *via* des initiatives telles FAANG (http://www.faang.org) ou AQUAFAANG (https://www.aqua-faang.eu/). Les populations animales présentent certains avantages importants pour les études de génétique systémique, comme la disponibilité de tissus pertinents et la capacité à contrôler l'environnement de ces études. Ces méthodes ont d'ores et déjà été appliquées et des résultats encourageants ont été obtenus chez le porc (Keel et al., 2020) et le bovin (Duarte et al., 2019) sur les caractères d'efficacité alimentaire. Chez le porc, Keel et al. ont développé une méthode permettant d'améliorer la puissance des analyses d'associations pangénomiques en attribuant un poids, évalué à partir de données transcriptomiques, aux variants identifiés (Keel et al., 2020). Chez le bovin, Duarte et al. ont identifié une voie biologique commune (valine, leucine and isoleucine degradation) associée à l'efficacité alimentaire à partir de plusieurs jeux de données génétiques obtenues dans différentes races bovines (Duarte et al., 2019). Cette même voie a également été mise en avant à partir d'analyses transcriptomiques entre 2 lignées bovines divergentes sur le critère de consommation résiduelle

alimentaire (Khansefid et al., 2017). Ces 2 exemples démontrent ainsi l'intérêt des approches de génétique systémique chez les animaux d'élevage pour décortiquer finement la relation genome-phenome *via* l'amélioration des connaissances sur l'architecture moléculaires des caractères agronomiques. Ainsi, l'utilisation de cette stratégie pour des nouveaux phénotypes d'intérêt comme la santé ou le bien-être contribuera à un élevage et des productions animales durables (Suravajhala et al., 2016).

Chapitre 2 : Identification de régions QTL pour la CMJR à partir des lignées divergentes INRAE

I. Introduction

Les deux lignées divergentes sur le CMJR (dispositif INRAE) obtenues après 10 générations de sélection sur une population de porcs pure race Large White, constituent un dispositif puissant pour l'étude de l'architecture génétique de ce caractère. Dans ce dispositif expérimental, l'ensemble des individus reproducteurs ont été génotypés avec une des puces MD permettant ainsi d'avoir des génotypes pour 60 000 SNPs (60K, Illumina Porcine SNP60v2 BeadChip) ou 70 000 SNPs (70K, Illumina Porcine HD Array GGP) du génome porcin. En complément de ces données génétiques MD, 32 fondateurs ont été génotypés à l'aide d'une puce HD. Ce chapitre est donc destiné à présenter le travail de cartographie génétique réalisé en combinant ces jeux de données génétiques à des mesures phénotypiques enregistrées à chaque génération. Si les animaux reproducteurs ont été phénotypés afin de calculer un index de sélection, la majorité des enregistrements (dont des mesures de qualité de la viande mesurées post-mortem) ont été obtenus sur des animaux issus à chaque génération de parités complémentaires destinées à évaluer la réponse à la sélection. Le dispositif était donc constitué de 1 632 animaux génotypés sur une puce MD, 32 fondateurs disposant de génotypes HD et 2 426 animaux réponses(P2) sans génotype mais phénotypés pour 24 caractères différents. Au vu de ces données, des étapes d'imputation ont été réalisées afin d'aboutir des génotypes HD pour l'ensemble des animaux reproducteurs. A partir de ces génotypes parentaux, des génotypes moyens ont été calculés pour l'ensemble des animaux réponses et des analyses GWAS ont été réalisées à partir de ces individus. Ces analyses ont été reportées dans un article soumis à GSE (Genetics Selection Evolution, BMC) le 28 octobre 2020. Les régions en ségrégation détectées ont alors été étudiées dans le but de comparer et comprendre l'impact de la sélection sur la CMJR sur le génome des porcs Large White sous sélection divergente.

II. Stratégie

Les différentes étapes de l'étude sont résumées dans la figure 33. Trois grandes étapes peuvent être distinguées, (*i*) le nettoyage des données, (*ii*) les imputations et (*iii*) les études GWAS.

Le nettoyage des données : Comme évoqué précédemment les données de génotypage disponibles ont été obtenues à des périodes différentes au cours des générations de sélection et 3 supports de génotypage différents ont été utilisés. Afin de disposer d'un jeu de données de bonne qualité, deux type de contrôle ont été réalisés, un contrôle qualité "classique" (filtrage à partir des "call-rate", "Call-freq", filtre sur la MAF et test d'équilibre de Hardy-Weinberg) puis un contrôle des données pedigree (contrôle de filiation, du sexe, de la génération et de la lignée) (Figure 33a).

Les imputations : Deux étapes d'imputations successives ont été réalisées. Entre les supports 60K (Illumina Porcine SNP60v2 BeadChip) et 70K (Illumina Porcine HD Array GGP), dont ¾ des variants sont communs aux deux puces MD. Une première étape d'imputation a été réalisée pour homogénéiser les données génétiques dans le but d'obtenir un seul jeu de données génétiques MD communes à tous les animaux sélectionnés (Figure 33b). Ce premier jeu de données de près de 60 000 SNPs (après filtrage et contrôle qualité) a été utilisé dans l'étude réalisée par Aliakbari et al. (Aliakbari et al., 2020) sur l'effet du degré d'apparentement entre individus issus des deux lignées dans le constitution d'une population de référence pour la prédiction génomique de la CMJR. Ce jeu de génotypes MD a également été utilisé dans une seconde étape d'imputation avec les génotypes HD disponibles pour 32 fondateurs, afin d'obtenir des génotypes HD pour l'ensemble des individus reproducteurs. Au total, plus de 550K marqueurs ont été imputés et ces génotypes ont alors été utilisés pour attribuer un génotype parental moyen aux animaux P2 (Figure 33b).

Les études GWAS : Les animaux P2 disposant alors des génotypes parentaux moyens, en plus de mesures phénotypiques pour plusieurs caractères d'intérêts (efficacité alimentaire, croissance et qualité de la viande après abattage), ont permis de mener des études GWAS pour identifier des régions génomiques en ségrégation (Figure 33c).

L'ensemble de ce travail est rapporté dans l'article présenté dans la troisième partie de ce chapitre. Au préalable j'ai néanmoins choisi de présenter l'étape de contrôle qualité du pedigree qui n'a pas été inclus à cet article.

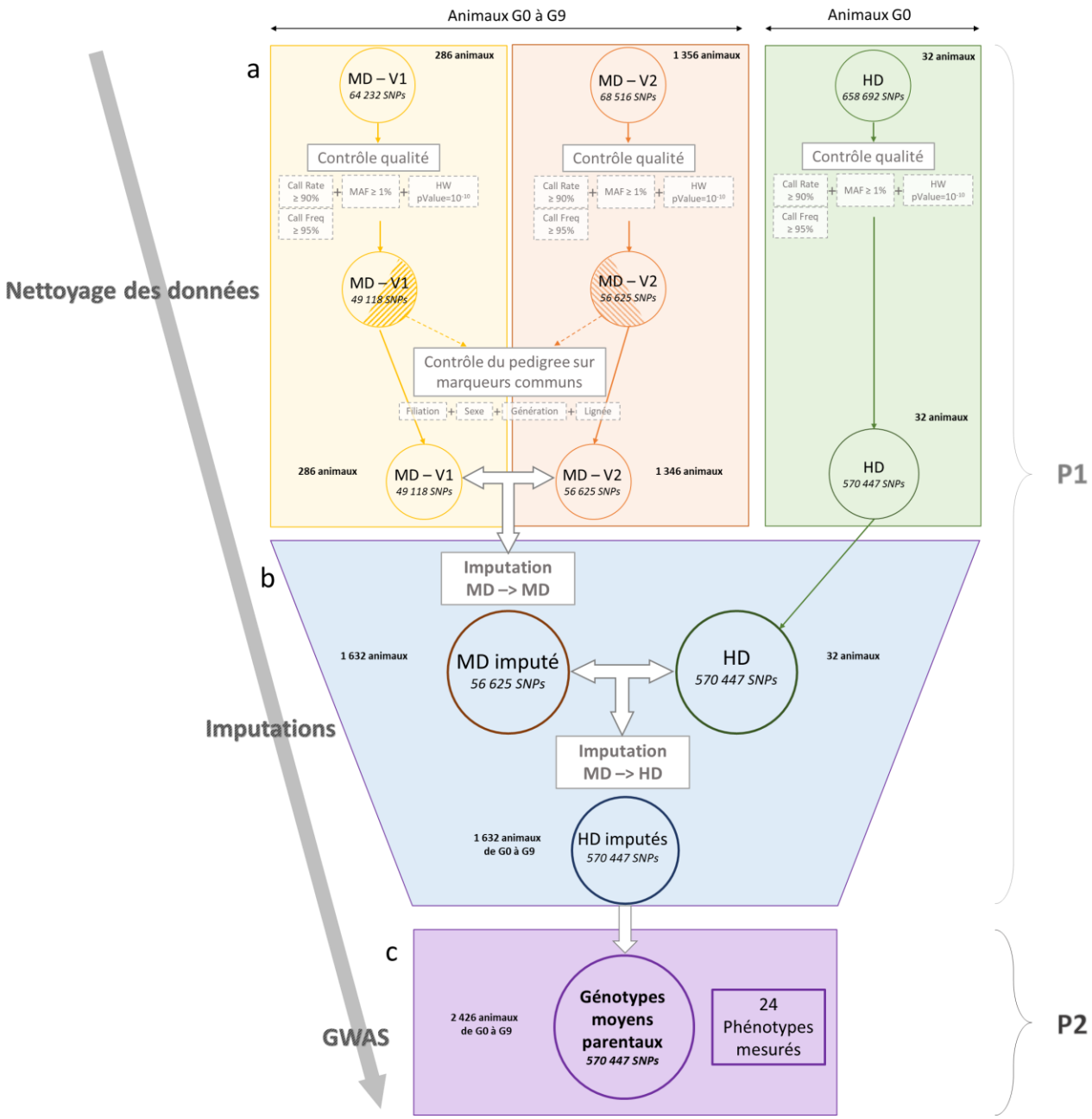

Figure 33 : Principe d'analyse des données génétiques issus du dispositif de sélection divergente pour la CMJR. (a) Trois puces génétiques ont été utilisées : une puce MD 60K (MD-V1, Illumina Porcine SNP60v2 BeadChip), une puce MD 70K (MD-V2, Illumina Porcine HD Array GGP) et une puce HD (Affymetrix Axiom Porcine HD Genotyping Array). Pour ces 3 jeux de données, un contrôle qualité sur les données génétiques a été appliqué et un contrôle du pedigree a également été réalisé grâce aux données génétiques MD. (b) Avec l'ensemble des données deux étapes d'imputation ont été réalisées, une première imputation pour homogénéiser les données MD et une seconde imputation pour avoir l'ensemble des individus avec des génotypes HD. Enfin un génotype parental moyen a été affecté aux animaux de la seconde parité grâce aux données génétiques imputés en HD pour les animaux P1. (c) Au final, des GWAS ont été réalisées avec les génotypes parentaux moyens et des phénotypes directement mesurées sur ces individus.

III. Identification et correction des erreurs de pedigree

Après avoir appliqué un contrôle qualité sur les individus et les SNPs, une étape de contrôle du pedigree où figurent chaque individu reproducteur du dispositif et les informations sur ses géniteurs, son sexe, sa génération et la lignée à laquelle il appartient, a été réalisé. Pour ces contrôles, seuls les 42 800 SNPs communs aux 2 puces MD ont été utilisés afin de pouvoir étudier l'ensemble des reproducteurs simultanément. Ces analyses avaient pour but d'obtenir un pedigree de qualité permettant de tirer un maximum profit du dispositif expérimental de sélection divergente des porcs Large White.

1. Contrôle des erreurs de parentés

Pour la détection des erreurs mendéliennes, les génotypes des descendants ont été confrontés aux génotypes de leurs parents. Au total, 1 589 individus ont pu être testé. Les génotypes des individus sont codés 0, 1 ou 2, 0 et 2 correspondant aux génotypes homozygotes pour l'allèle mineur ou l'allèle majeur et 1 correspondant au génotype hétérozygote. Sachant qu'un individu récupère à la fois de l'information génétique de son père et de sa mère, plusieurs génotypes finaux sont possibles lors du croisement et toutes les combinaisons possibles sont identifiées dans la figure 34.

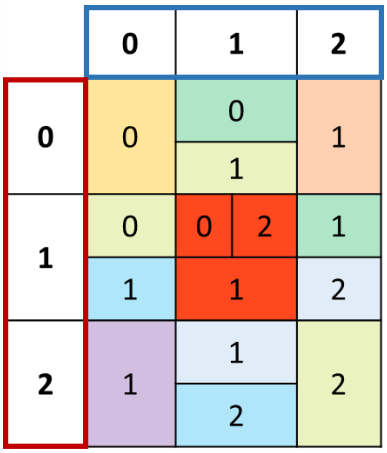

Figure 34 : Différents génotypes de descendants lors du croisement d'un mâle (génotypes 0/1/2 encadré bleu) avec une femelle (génotypes 0/1/2 encadré rouge) sont possibles. Néanmoins, si les deux parents sont codés 1, tous les génotypes sont possibles (cases rouges).

Pour la détection des erreurs de parenté, les SNPs pour lesquels les deux parents sont codés hétérozygotes (soit 1) n'ont pas été pris en compte car tous les génotypes sont possibles pour les descendants. Aucune erreur ne peut être détectée avec ce type de croisement. Pour toutes les autres situations nous avons appliqué le calcul suivant :

$$|(Genotype_{descendant} - Genotype_{père}) + (Genotype_{descendant} - Genotype_{mère})|$$

Si le score obtenu est inférieur ou égal à 1, le génotype testé de l'individu est alors vérifié et validé. En revanche si le score est strictement supérieur à 1, cela signifie que le génotype du descendant n'est pas possible au vu des génotypes parentaux. Si on prend pour exemple un descendant hétérozygote

(codé 1) alors que ces deux parents sont homozygotes pour l'allèle mineur (codé 0) le score obtenu sera de 2 (|(1-0)+(1-0)| = 2). Ce génotype n'est évidemment pas possible. En revanche si l'un de ces deux parents est hétérozygote (codé 1), cette fois le résultat obtenu est de 1 (|(1-0)+(1-1)|=1) et le génotype est alors validé. Pour chaque trio père, mère, descendant, les scores ont été calculés pour chaque SNP informatif et le nombre de SNP dont la valeur obtenue indiquait une erreur mendélienne a été comptabilisé. La majorité des individus présentaient au maximum 300 erreurs, correspondant à un taux d'erreur de génotypage de moins de 1%. Néanmoins, 54 individus ont présenté plus de 1 000 SNP avec des erreurs mendéliennes (Figure 35).

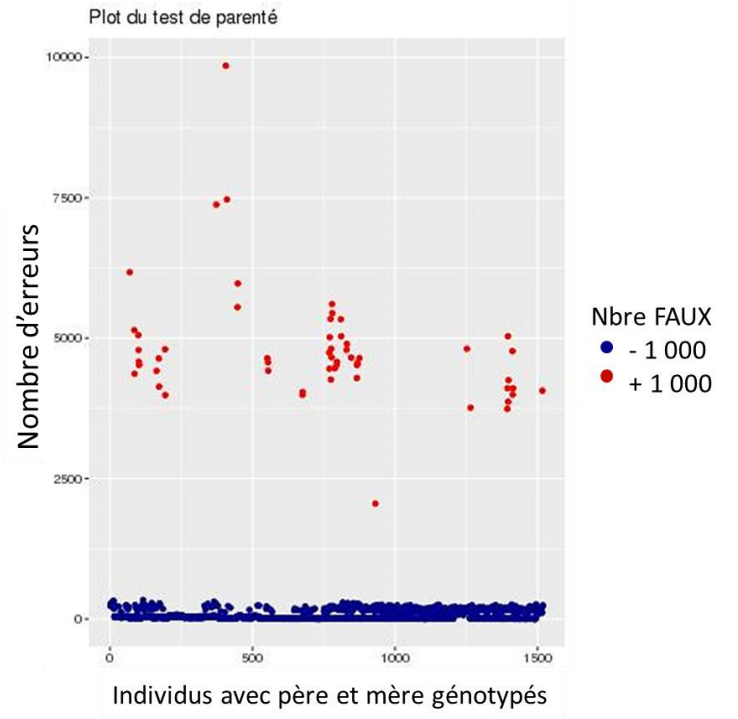

Figure 35 : Résultat du test de parenté pour les 1 589 individus dont les deux parents ont été génotypés. Les individus bleus ne possèdent que quelques erreurs mendéliennes pour les 42 800 SNPs testés, en revanche les individus rouges sont détectés à problème car plus de 1 000 erreurs mendéliennes ont été détectées.

Pour les 54 individus présentant des erreurs mendéliennes nous avons alors recherché dans l'ensemble du pedigree les vrais couples père-mère avec lesquels leurs génotypes seraient compatibles. Pour cela, le calcul présenté précédemment a été utilisé afin de tester toutes les combinaisons de parents possibles pour ces 54 individus à erreur de parenté. Pour 39 animaux nous avons ainsi pu réattribuer leurs vrais parents. Les 15 individus restant étaient à la fois incompatibles avec leurs deux parents mais également comme parents avec l'ensemble de leurs descendants. L'explication la plus probable est que l'échantillon d'ADN génotypé ne correspondait pas à l'animal supposé. Nous avons donc fait le choix de supprimer leurs génotypes afin de les "reconstruire" à partir des génotypes de leurs descendances. Cette approche nous a permis d'obtenir un génotype compatible et utilisable pour les analyses suivantes pour 6 individus supplémentaires (1 verrat et 5

truies disposant d'au moins 4 descendants). Au final, 9 femelles ont dû être supprimées du jeu de données car le nombre de descendant n'était pas assez important pour imputer leurs génotypes.

2. Les erreurs de sexe

Afin de contrôler le sexe des individus indiqué dans le fichier généalogique, j'ai sélectionné les génotypes des SNPs localisés sur le chromosome X. Les chromosomes sexuels présentent 2 portions caractéristiques (Figure 36). (*i*) Des portions homologues aussi appelées régions pseudo-autosomiques (PAR) qui permettent aux deux homologues de se reconnaître et de s'apparier à la méiose. Les SNP positionnés dans ces régions sont présents sur les chromosomes X et Y, et pourront présenter des génotypes 0, 1 ou 2 quel que soit le sexe de l'individu analysé (mâle et femelle). En revanche, (*ii*) la portion spécifique des chromosomes sexuels est spécifique à chacun des sexes, de ce fait les SNPs de la région spécifique du chromosome X sont uniquement présents sur ce chromosome. Les mâles ne portant qu'un seul chromosome X, aucun ne peut être hétérozygote pour les SNPs localisés dans cette portion de l'X.

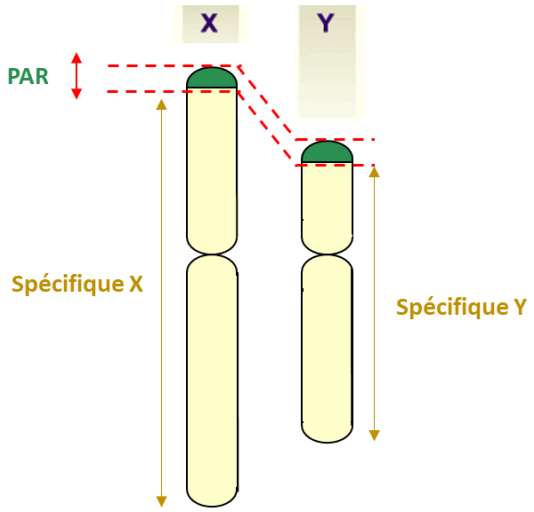

Figure 36 : Schéma des chromosomes X et Y qui possèdent une portion chromosomique homologue entre les deux chromosomes sexuels X et Y (PAR en vert) et une portion spécifique à chacun des chromosomes (spécifique X/Y en beige).

Le calcul du taux d'hétérozygotie pour chaque marqueur, à partir des mâles du dispositif, permet d'identifier facilement la limite entre les deux portions chromosomiques du X. Les marqueurs présentant un taux supérieur à 0 correspondent aux SNPs localisés sur la PAR, les marqueurs dont le taux est nul sont localisés dans la portion spécifique du X. Le taux d'hétérozygotie pour l'ensemble des SNPs du chromosome X en fonction de leurs positions sur la carte chromosomique est représenté sur la figure 37. Trois SNPs de la puce 60K présentaient une valeur supérieure à 0,1. Nous avons choisi d'exclure de l'analyse le marqueur présentant un taux d'hétérozygotie de 1 et de mettre en valeur manquante les génotypes obtenus pour les deux autres marqueurs, cette information manquante pouvant être comblée lors des étapes ultérieures d'imputation.

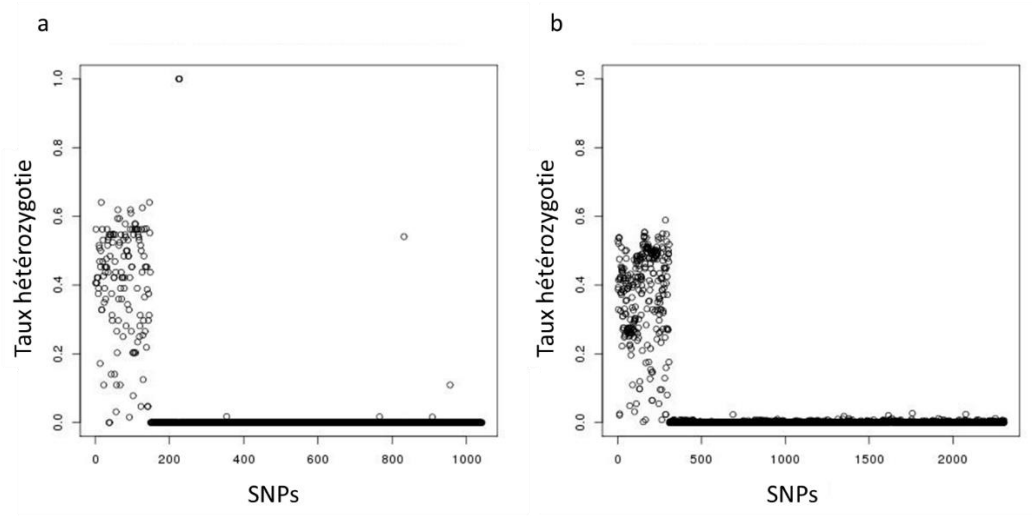

Figure 37 : Représentation des valeurs d'hétérozygotie des variants du chromosome X avec l'identification de la PAR et de la portion spécifique pour les mâles génotypés (a) sur la puce 60K et (b) sur la puce 70K.

Pour détecter des erreurs de sexe au sein du pedigree, j'ai alors calculé le taux d'hétérozygotie de l'ensemble des individus génotypés à partir de l'ensemble des SNPs de la région spécifique du chromosome X. Ainsi, tous les mâles doivent obligatoirement être homozygotes pour l'ensemble des marqueurs (Figure 38). Au total, cette approche nous a permis d'identifier deux individus à problème dont le sexe a été modifié dans le fichier pedigree.

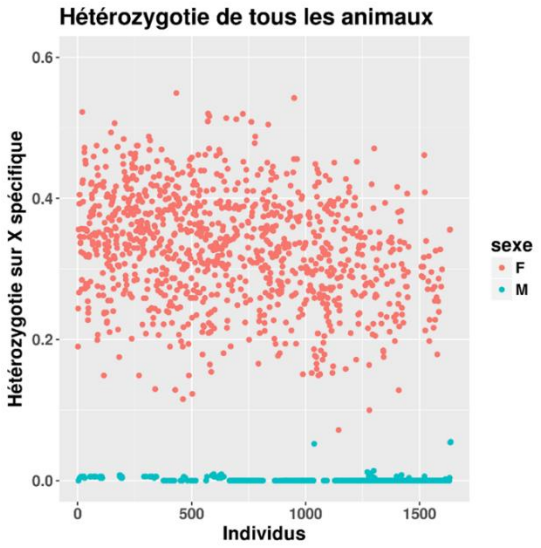

Figure 38 : Représentation de l'hétérozygotie des individus génotypés avec la distinction des femelles (rouge) et des mâles (bleu) au vu des valeurs de fréquences alléliques

3. Les erreurs de générations et de lignées

Les derniers contrôles réalisés ont porté sur le numéro de génération et l'appartenance à une des deux lignées. Au cours des générations, différents fichiers de pedigree se sont ajoutés et des divergences entre fichiers ou des données manquantes ont pu être corrigées au sein du pedigree. Les erreurs de générations et de lignées ont été détectées à l'aide des géniteurs. En effet tous les descendants

devaient appartenir à la même lignée que ses parents et les 2 parents devaient eux aussi être issus de la même lignée. De plus, chaque descendant devait être rattaché à la génération N+1 de ses parents. Grâce à ces contrôles, des données manquantes ont pu être complétées et deux erreurs de lignées corrigées.

Une fois les données du pedigree corrigées et validées, une première étape d'imputation a pu être réalisée en utilisant le logiciel FImpute qui permettait de prendre en compte les données familiales pour améliorer la qualité de l'imputation. Les 1 632 animaux possédant ainsi un génotype MD imputé ont pu être utilisé dans l'étude menée par Aliakbari et al. (Aliakbari et al., 2020) **sur la prédiction génomique des porcs Large White pour la CMJR (Annexe 3).**

IV. Article : Identification des régions génomiques affectant les caractères de production des porcs sélectionnés de manière divergente pour l'efficacité alimentaire

Cet article rapporte la création du jeu de données utilisé pour les études GWAS à l'aide d'étapes successives d'imputation puis les différentes études d'associations réalisées pour 24 caractères. Des GWAS ont été réalisés à partir de la totalité des animaux du dispositif (Global-GWAS) ou en utilisant uniquement les individus d'une seule des deux lignée (HRFI-GWAS ou LRFI-GWAS). Peu de régions QTL communes entre toutes les analyses ont été identifiées, et afin de comprendre les causes de ces différences nous avons ensuite étudié l'évolution des fréquences alléliques au cours des générations dans les régions QTL identifiées. Enfin une analyse bibliographique des gènes présents dans ces régions a été réalisée pour identifier des gènes candidats fonctionnels et positionnels pour ces QTL.

Identification of genomic regions affecting production traits in pigs divergently selected for feed efficiency

Emilie Delpuech[1], Amir Aliakbari[1], Yann Labrune[1], Katia Fève[1], Yvon Billon[2], Hélène Gilbert[1], Juliette Riquet[1]*

[1]GenPhySE, Université de Toulouse, INRAE, F-31326 Castanet-Tolosan, France

[2]GenESI, INRAE, F-17700 Surgères, France

Abstract

Background

Feed efficiency is a major driver of the sustainability of pig production systems. Understanding biological mechanisms underlying these agronomic traits is an important issue whether for environment and farms economy. This study aimed at identifying genomic regions affecting residual feed intake (RFI) and other production traits in two pig lines divergently selected for RFI during 9 generations (LRFI, low RFI; HRFI, high RFI).

Results

We built a whole dataset of 570,447 single nucleotide polymorphisms (SNPs) in 2,426 pigs with records for 24 production traits after both imputation and prediction of genotypes using pedigree information. Genome-wide association studies (GWAS) were performed including both lines (Global-GWAS) or each line independently (LRFI-GWAS and HRFI-GWAS). A total of 54 chromosomic regions were detected with the Global-GWAS, whereas 37 and 61 regions were detected in LRFI-GWAS and HRFI-GWAS, respectively. Among those, only 15 regions were shared between at least two analyses, and only one was common between the three GWAS but affecting different traits. Among the 12 QTL detected for RFI, some were close to QTL detected for meat quality traits and 9 pinpointed novel genomic regions for some harbored candidate genes involved in cell proliferation and differentiation processes of gastrointestinal tissues or lipid metabolism-related signaling pathways. Detection of mostly different QTL regions between the three designs suggests the strong impact of the dataset on the detection power, which could be due to the changes of allelic frequencies during the line selection.

Conclusions

Besides efficiently detecting known and new QTL regions for feed efficiency, the combination of GWAS carried out per line or simultaneously using all individuals highlighted the identification of chromosomic regions under selection that affect various production traits.

Background

Feed efficiency is a major driver of the sustainability of pig production systems. It represents from 50 to 83 % of cost productions depending on the countries and systems [1]. The feed efficiency is also a principal lever to reduce the environmental footprints of the production [2]. The cost of feeding in pig production is usually measured by the computation of feed conversion ratio (FCR). Indeed, FCR is a ratio between two traits of interest for most breeding schemes (feed intake and growth rate), and

incorporating it in selection indexes makes it difficult to accurately anticipate responses to selection on this trait and the correlated traits [3]. In 1963, Koch et al. [4] proposed residual feed intake (RFI) as an alternative to quantify feed efficiency, to overcome the limits of FCR. The RFI is the difference between individual feed intakes and predicted feed intake for the animal maintenance and production requirements. It is generally computed as a multiple linear regression of daily feed intake on production traits (growth rate and body composition traits in growing animals), and on the average metabolic body weight of the animal during the period, as an indicator of maintenance requirements. As a result, selection for RFI generates limited correlated responses on the other production traits, as shown in several selection experiments in pigs [5, 6], and other species [7]. However, accurate individual feed intake recording for pigs raised in groups is costly, and large efforts are devoted to facilitate the improvement of feed efficiency, by either identifying biomarkers [8, 9] or genomic markers (for instance [10, 11]). Despite these efforts, the difficulty to find quantitative trait loci (QTL) or genomic variant affecting feed efficiency related traits translates in the PigQTLDB statistics [12]: only 394 QTL are listed for feed conversion type of traits, and 350 for feed intake type of traits, whereas more than 2,000 are reported for growth traits, and more than 3,200 for fatness traits (PigQTLDB, access Sept 2020, https://www.animalgenome.org/cgi-bin/QTLdb/SS/index). Genomic information acquired in established divergent lines for the trait of interest can be used to increase the power of detection of genomic variants for lowly heritable or highly polygenic traits, as RFI in pigs [10] and litter traits in rabbits [13].

In this study, we aimed at identifying genomic regions affecting RFI and other production traits in two pig lines divergently selected for RFI during 9 generations [5], by combining an extensive genotyping of all breeding animals of the lines, and the extensive phenotyping of their progeny. GWAS were applied to growth, feed intake and feed efficiency, carcass composition and meat quality traits on the full dataset. Different subsets of the population were used to suggest biological hypotheses for the genetic background of the traits in the two divergent lines, and decipher whether the chromosomic regions affecting RFI differed between lines.

Methods

Ethic statement

All pigs were reared in compliance with national regulations and according to procedures approved by the French Veterinary Services at INRA experimental facilities. The care and use of pigs were performed following the guidelines edited by the French Ministries of High Education, Research and Innovation, and of Agriculture and Food (http://ethique.ipbs.fr/sdv/charteexpeanimale.pdf).

Design

The data were obtained from a divergent selection experiment on RFI carried out at the INRA experimental units GenESI since 2000 (Surgères, France, https://doi.org/10.15454/1.5572415481185847E12), on growing pigs from the French Large-White (LW) population. The selection procedures were described by Gilbert et al. [5]. In brief, the lines were established from 30 matings of LW animals (F0). From these litters, 116 males were tested to select the 6 most efficient (LRFI) and 6 least efficient (HRFI) males as founders of two divergent lines, and about 40 pairs of sibs were randomly assigned to each line. In the following generations, from G1 to G9, 96 males from each line were tested for RFI to select 6 extreme low or high boars depending on the line. In addition, 35 to 40 females were randomly chosen within-line in each generation to produce the next generation. No selection was applied for females. From G1, matings were organized for at least two successive litters. Until G5, the first litter provided boars candidates for selection and future breeding females, and castrated males and females from the second parity were tested to evaluate the direct and correlated responses to selection on major production traits, including carcass composition and meat quality traits. After G5, selection was applied to parity 4 or 5, and responses to selection were measured on pigs born in parity 2 and 3. Hereafter, the breeding animals will be called "breeders" and animals tested for responses to selection will be called "response animals".

Phenotypes

In each generation, 48 females and 48 castrated males per line were produced as response animals, and tested individually during the growing-finishing period (~28 kg to ~107 kg) for body weight (BW0 at the start of the test and BW1 before slaughter) and daily feed intake (DFI) using a single-place electronic feeder (ACEMA 64; Skiold Acemo, Pontivy, France) to compute average daily gain (ADG) and feed conversion ratio (FCR) during the test period. The dressing percentage (DP) was computed based on weight records of warm carcass at slaughter. Twenty four hours after slaughter, backfat thickness measured on carcass (carcBFT), and the weights of ham (Ham_W), loin (Loin_W), belly (Belly_W), shoulder (Shoulder_W), and backfat (BF_W), following a standardized cut, were recorded on the cold half carcass. The lean meat content (LMCcalc) was evaluated according to the method of Daumas [14]. Meat quality measurements included pH on *adductor femoris* muscle (AD), *semimembranosus* muscle (SM), *gluteus superficialis* muscle (GS), and *longissimus dorsi* muscle (LM), colorimetry L*, a* and b* on GS and *gluteus medius* muscle (GM), and water-holding capacity (WHC) assessed on GS according to the procedure described by Charpentier et al. [15]. Finally, meat quality index (MQI) was calculated from measurements of pH in SM, L* on GS and WHC according to the model proposed by Tribout et al. [16]. RFI was defined as the residual of a multiple linear regression as follows: RFI = DFI – (1.48 ×

ADG) + (23.2 × LMCcalc) – (99.1 × AMBW), where AMBW is the average metabolic body weight during the test period and is equal to $(BW1^{1.6} - BW0^{1.6})/[1.6\ (BW1 - BW0)]$ [17]. Contemporary group, gender and pen size were added as fixed effects in the model, as described by Gilbert et al. [5].

Genotyping

Genomic DNA was purified from individual biological samples of the sires and dams of all generations using standard protocols. Over time, two different Illumina medium density SNPs chips were used according to the genotyping protocols defined by the supplier (at Technological Center, Genomics and Transcriptomics Platform, CRCT Toulouse). First batch comprising 286 animals was genotyped for 64,232 SNPs using the Porcine SNP60v2 BeadChip (60K SNPs chip), and a second batch of 1,356 animals was genotyped using the Porcine HD Array GGP chip comprising 68,516 SNPs (70K SNPs chip). Genotypes were obtained using the Genome Studio software (V2.0.4) and coded as 0, 1 and 2 corresponding, respectively, to individuals homozygous for the minor allele, heterozygous and homozygous for the major allele. In addition, 32 G0 founders (12 G0 sires, and 20 G0 dams that had most contribution to the subsequent generations) were genotyped with the Affymetrix Axiom Porcine HD Genotyping Array chip (Gentyane Plateform, UMR 1095 INRAE Clermont-Ferrand) consisting of 658,692 SNPs (650K SNPs chip).

For each SNPs panel, quality control was performed using PLINK software (V1.90) [18]: SNPs with a call frequency (CF) < 95% and a minor allele frequency (MAF) < 1% were excluded, and animals with a call rate (CR) <90% were discarded. Unmapped SNPs and SNPs located on sex chromosomes were removed following the Sscrofa11.1 assembly of the reference genome (https://www.ensembl.org/Sus_scrofa/Info/Index)[19].

Genotypes imputation

Two successive imputations were performed using the FImpute software [20]. A first level of imputation was performed with markers of 60K and 70K SNPs chips, based on 29,957 SNPs in common, to homogenize the medium density genotyping data available for the 1,632 breeders of the lines. This leads to an intermediate dataset of 66,988 SNPs imputed from both medium density (MD) chips (60K and 70K SNPs chips). In a second step, the genotypes of the high density (HD) SNPs chip were imputed for all breeders using the HD SNPs genotypes of the 32 G0 founders. A set of 45,708 SNPs was in common between MD imputed genotypes and HD SNPs chip. A total of 570,447 SNPs distributed over the 18 pig autosomes, was finally available for 1,632 breeding animals.

To evaluate the imputation accuracy, first, five successive batches of 1,000 SNPs were randomly selected among the common SNPs in the 60K and 70K SNPs chips. For each SNPs batch, the

genotypes of these SNPs were set as missing for all animals genotyped with the 60K SNPs chip and imputed from the 70K SNPs chip information. Therefore, a total of 5,000 SNPs with real and imputed genotypes were used to compute Pearson correlations for each of the 286 pigs with 60K genotypes. Similarly, five batches of 1,000 SNPs were randomly selected from common markers of both MD SNPs chips, animals genotyped with the 70K SNPs support were re-coded as missing, and Pearson correlations between true and imputed genotypes were computed for the 1,346 animals with 70K SNPs genotypes. Then, to evaluate the imputation quality to the HD, the same strategy of removing successively five batches of 1,000 SNPs from the data was applied using SNPs in common to the three chips. In addition, a leave-one-out approach was applied to the 32 individuals with HD genotypes to evaluate the imputation accuracy.

In addition, a multi-dimensional scaling (MDS) analysis was performed using R software (V.3.6.2, R Core Team 2019) based on a identity-by-state matrix constructed with the PLINK software [21].

Predicted genotypes in response animals

Response animals did not have genotypes themselves. An average expected genotypes of their parents was computed for each animal from the imputed 650K genotypes. For each marker, each individual was given the average genotype of the parents (0, 0.5, 1, 1.5 or 2), so within a litter, all animals were assigned the same genotypes. Depending on the genotypic class, the obtained genotype was, therefore, an approximation of the real genotype: (*i*) genotypes 0 and 2 were certain, as they resulted from two homozygous parents for the same allele (0x0 → 0 and 2x2 → 2), (*ii*) genotypes 0.5 and 1.5 included combinations of a homozygous genotype for one allele and a heterozygous genotype (0x1 → 0 or 1 and 1x2 → 1 or 2), and (*iii*) genotype 1 was the most heterogeneous class, with a mixture of true genotypes (0x2 → 1) and uncertain genotypes (1x1 → 0 or 1 or 2). Animals with a parent with a missing genotype were excluded from the analysis.

Genome-Wide Association Studies

GWAS analyses were performed using GEMMA software (version 0.97) [22] on response animals with their own phenotypes and their average genotypes from parents. Phenotypes were adjusted for significant fixed effects and covariates (pen size, herd, sex, and contemporary groups for *in vivo* measurements, slaughter date as fixed effects, and slaughter age as covariate for traits recorded at the abattoir, and slaughter BW as covariate for carcBFT) using linear models as proposed in Aliakbari et al. [23]. The resulting residues were integrated as phenotypes in GEMMA. To account for the structure of

the population in the GWAS analyses, a pedigree relationship matrix **A** was computed. Association analyses were performed on the 24 traits available for 2,426 response animals.

The statistical model used to test one marker at a time was $\mathbf{y} = \mathbf{x}\beta + \mathbf{u} + \boldsymbol{\varepsilon}$, where $\mathbf{y}$ is the vector of adjusted phenotypes for all individuals; $\mathbf{x}$ is a vector of genotypes at the tested marker; β is the effect of the tested marker; $\mathbf{u}$ is a vector of random additive genetic effects distributed according to $N(0, \mathbf{A}\lambda\tau^{-1})$, with λ the ratio of the additive genetic variance and the residual variance τ^{-1}; $\boldsymbol{\varepsilon}$ is a vector of residuals $N(0, \mathbf{I}\tau^{-1})$, with **I** the identity matrix. In GEMMA, an efficient exact algorithm is implemented to first estimate λ, and next derive $\hat{\beta}$ and $\hat{\tau}$ for each marker [24].

The distributions of the *p*-values for the GWAS of each trait were checked using quantile-quantile plots (Q-Q plot) and computing regression coefficients of the -$\log_{10}$(observed *p*-values) on the -$\log_{10}$(expected *p*-values under H_0). Inflation factors were lower than 1.23 for all analyses, indicating low deviations from the distribution of the test statistic under H_0. A correction factor was anyway applied to all analyses to control type-I errors, by dividing each *p*-value by the corresponding inflation factor to avoid the impact of this low deviation.

To account for the nominative type-I error and multiple testing issue, the significance threshold was obtained after a Bonferroni correction as follows:

$$-log10\left(\frac{0.05}{\sum(number\ of\ independent\ tests/chromosome)}\right),$$

where number of independent tests was computed per chromosome as the number of principal components required to describe 99.5% of the genotypic variability of each chromosome, after a principal component analysis applied to the correlation matrix between genotypes of the SNPs of the considered chromosome (square root (r^2) of linkage disequilibrium (LD) between each pair of SNPs), as recommended by Gao et al. [25]. The chromosome-wide thresholds obtained were between 3.09 and 3.16, thus a threshold of 3 was used to identify suggestive associations. To determine genome-wide significance thresholds, the number of independent tests for the 18 autosomes was summed to apply a correction at the genome level. This threshold (4.5) was used to identify significant associations.

Three types of populations were considered for GWAS. First, the full dataset combining the two lines was analyzed in a global analysis (thereafter called Global-GWAS). Then, to evaluate if some QTL were segregating in one line only, the analyses were repeated within line (thereafter called Lines-GWAS, or HRFI-GWAS and LRFI-GWAS when only one line was referred too).

To define QTL intervals, for each combination of population and trait, the genome was divided into 1 Mb windows following the Sscrofa11.1 assembly of the reference genome. The 1 Mb windows with at least one SNP with significant *p*-value at 5% genome-wide (-$\log_{10}$(*p*-value)≥4.5) were retained, and adjacent windows with significant signals were combined into a single QTL window per trait. In a second step, the adjacent and overlapping significant windows between traits were combined using

the same approach as presented above, thus allowing a complete list of QTL regions to be subsequently analyzed. When a QTL region was significant for several traits, for each of them, the most significant marker and the associated allelic substitution effect was retained to tag the QTL (trait x region) for this trait in further analyses – thereafter called SNP-QTL.

The QTL positions were compared to previously mapped QTL in pigs using the pigQTLdb database [12], and QTL significant for RFI trait were screened for functional candidate genes using Ensembl annotation V.101 (August 2020).

Changes of allelic frequencies of SNP-QTL

The power of detection in GWAS is strongly influenced by the allelic frequencies of the analyzed markers [26]. Within each QTL window, the most significant SNP was considered to examine the changes of allele frequencies with line selection. These SNP-QTL allele frequencies were estimated for the response animal genotypes, i.e. from their average genotypes. To find out how selection affected allele frequencies, and thus power of detection, allele frequencies were computed by adding animals from one generation at a time, starting from G1 individuals alone. Then, the allele frequencies adding G2 response animals were obtained by combining genotypes of G1 and G2 response animals, and so on until G9. The estimated frequencies in G9 (using all the animals from G1 to G9) corresponded to the informativeness of the markers used in the main GWAS by line. A regression of the generation (1 to 9) on the SNP allele frequencies was then applied to test changes of allelic frequencies on cumulative datasets over generations. For each SNP-QTL, the significance of the slope was estimated in each line using a Wald test. An average evolution score was compute for each QTL region (9 generations * ($|slope_{HRFI}| + |slope_{LRFI}|$)) when the slope value was different from zero with $p < 0.05$. To reflect the evolution per trait, an average value over all the QTL regions detected for each trait was computed.

Results

Genotype quality control and imputation

True SNPs genotyping data were available for all sires and dams from G0 to G9. The quality control of the genotypes was first carried out for each SNP chip independently. With a CR threshold of 90%, 10 animals genotyped with the 70k SNPs chip and no individual genotyped with the 60K and 650K SNPs chips were discarded (Additional file 1). For the SNPs, 15,114 SNPs from the 60K SNPs chip (5,776 for CF < 95% and 9,125 for MAF < 1%), 11,891 SNPs from the 70K SNPs chip (5,323 for CF < 95% and 6,568 for MAF < 1%), and 99,587 SNPs from the HD SNPs chip (53,735 for CF < 95% and 45,852 for MAF < 1%) were removed. In total, genotypes of 286 animals for 49,118 SNPs for the 60k SNPs chip, genotypes

for 1,346 animals for 56,625 SNPs for the 70K SNPs chip, and finally genotypes for 32 animals for 559,105 SNPs for the HD SNPs chip were retained for further analyses (Additional file 2).

To obtain HD genotypes for all parents of the design, two successive runs of imputations were performed. First, the imputation of the missing genotypes on each MD support (60K and 70K SNPs chips) allowed obtaining genotypes for 66,988 SNPs for all sires and dams. The imputation accuracy was on average 0.995 regardless the generation of the imputed individuals (Figures 1a and 1b). A second run of imputation was applied to all breeding animals from the 32 founder individuals genotyped with the HD SNPs chip. The imputation accuracy was also high, with average accuracies around 0.979 (Figure 1c). Some few animals in G0 and G3 had accuracies lower than 0.97. The accuracy estimated via the leave-one-out approach confirmed the values estimated with the correlations, with an average of 0.975. In total, genotypes for 570,447 SNPs were obtained for all parents from G0 to G9.

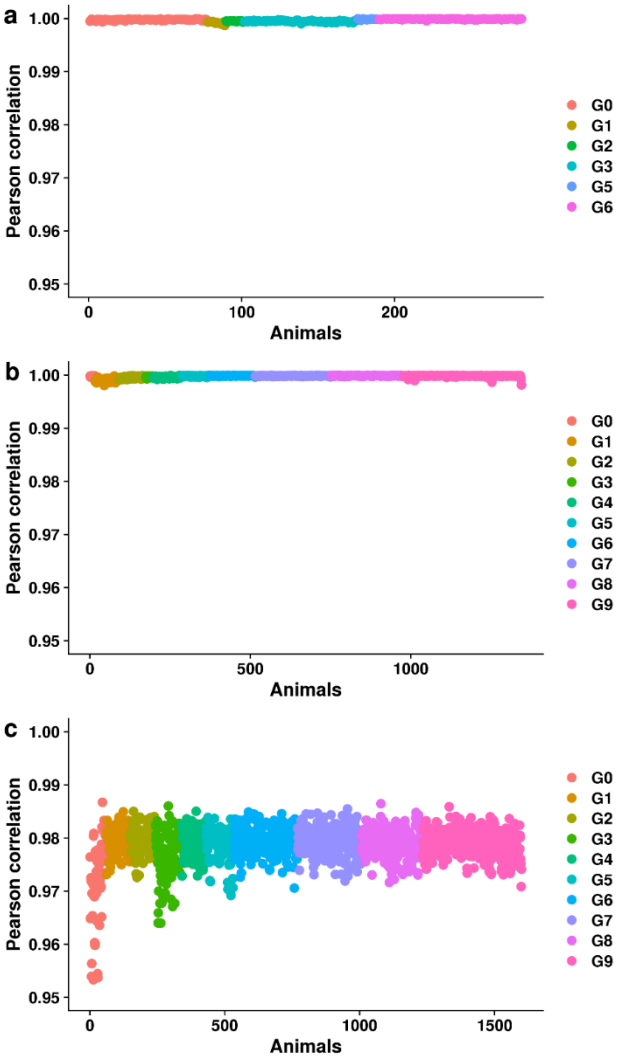

Figure 1 Correlations between true and imputed genotypes for animals genotyped on 60K or 70K SNPs chip. For each analysis, correlations were estimated setting 5,000 SNPs as missing (5 batches of 1,000 SNPs) on one chip among SNPs in common between the two supports used. Animals are sorted and colored by generation. Correlations between true and imputed genotypes (a) for the 286 animals genotyped with the 60K SNPs chip using animals with 70K genotypes as reference

population, and (b) for the 1,346 animals genotyped with the 70K SNPs chip using animals with 60K genotypes as reference. (c) Correlations between true and imputed genotypes after imputation to 650K SNPs from the imputed medium density genotypes.

An MDS analysis was performed on the genotypic matrix to represent the changes of genomic content of the lines with generations (Figure 2). The first component corresponded to the dispersion of individuals according to the lines, and the second component corresponded to the successive generations in both lines.

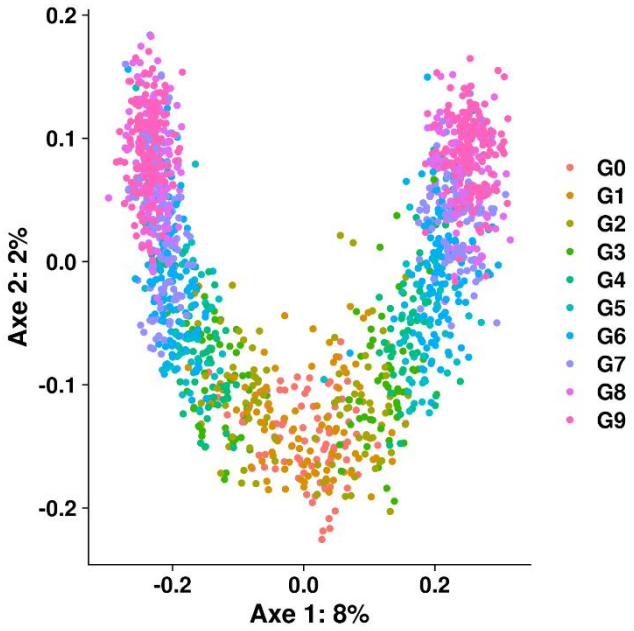

Figure 2 Two first axes of the multidimensional scaling (MDS) analysis, based on the 570,447 genotypes. Points represent individuals (corresponding to all sires and dams of the population, N=1,632) and colors are generations.

Genome-wide association studies

From the imputed genotypes of all parents, an average genotype was computed for all response animals. Thus genotypes coded 0, 0.5, 1, 1.5 or 2 were available for 2,426 individuals in total. Within a sibling, all individuals shared the same average genotype. On average the size of the siblings was 4.07 (± 2.9).

First, association studies corresponding to Global-GWAS were carried out on all response animals, for each of the 24 traits. A total of 54 regions of 1 Mb (38 regions), 2 Mb (12 regions), or 3 Mb (4 regions) were significant for at least one trait, corresponding to 72 QTL (trait x region). QTL were detected for all 24 traits (Figure 3), the list and characteristics of these QTL is reported in the Additional file 3. Cut weights were the traits with the lower number of QTL (1 to 3 per analysis), except for the weight of backfat (BF_W) in the Global-GWAS (Table 1). Meat quality measurements had the highest number of QTL (up to 7). Thirty regions associated with growth, feed intake, and feed efficiency were

detected, including 12 regions associated with RFI and 5 with FCR. For all traits (except Belly_W), at least one QTL was detected in the Global-GWAS.

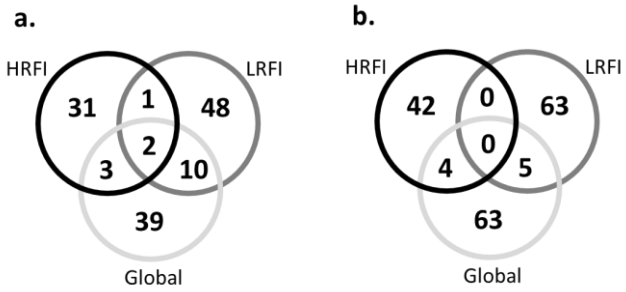

Figure 3 Location of all SNP-QTL identified on the 18 autosomes from the Global-GWAS, LRFI-GWAS and HRFI-GWAS. The SNP-QTL corresponding to Global-GWAS are represented by horizontal bars, LRFI-GWAS by arrows to the right of the chromosomes and HRFI-GWAS by arrows to the left of the chromosomes. Each color represents one of the 24 traits

LRFI: low RFI line, *HRFI*: high RFI line

DFI: daily feed intake*; ADG:* average daily gain*; FCR:* feed conversion ratio*; RFI:* residual feed intake*; carcBFT:* backfat thickness measured on carcass*; a*_GM:* a* measured on the *gluteus medius* muscle*; a*_GS:* a* measured on the *gluteus superficialis* muscle*; b*_GM:* b* measured on the *gluteus medius* muscle*; b*_GS:* b* measured on the *gluteus superficialis* muscle*; L*_GM:* L* measured on the *gluteus medius* muscle*; L*_GS:* L* measured on the *gluteus superficialis* muscle*; pH24h_AD:* pH 24h after slaughter measured on the adductor femoris muscle*; pH24h_GS:* pH 24h after slaughter measured on the *gluteus superficialis* muscle*; pH24h_LM:* pH 24h after slaughter measured on the *longissimus dorsi* muscle*; pH24h_SM:* pH 24h after slaughter measured on the *semimembranosus* muscle*; WHC:* water holding capacity of the *gluteus superficialis* muscle*; MQI:* meat quality index*; LMCcalc:* lean meat content of the carcass*; DP:* carcass dressing percentage*; Belly_W:* belly weight*; BF_W:* backfat weight*; Ham_W:* ham weight*; Loin_W:* loin weight*; Shoulder_W:* shoulder weight

Trait	Global	HRFI	LRFI	Total
DFI	3	2	4	9
ADG	1	3	0	4
FCR	2	1	2	5
RFI	3	0	9	12
carcBFT	4	1	5	10
a*_GM	4	0	1	5
a*_GS	1	4	4	9
b*_GM	3	1	1	5
b*_GS	7	5	1	13
L*_GM	1	5	2	8
L*_GS	3	4	6	13
pH24h_AD	2	0	3	5
pH24h_GS	5	1	2	8
pH24h_LM	4	3	6	13
pH24h_SM	4	1	2	7

WHC	3	6	1	10
MQI	4	2	2	8
LMCcalc	4	1	2	7
DP	3	1	7	11
Belly_W	0	0	2	2
BF_W	5	2	3	10
Ham_W	3	1	1	5
Loin_W	2	1	1	4
Shoulder_W	1	1	1	3
Total	59	39	48	186

Table 1 Number of QTL identified for each trait with the 3 groups of association studies. Association studies on the full population (Global-GWAS, *Global*) and for each line separately (HRFI-GWAS, *HRFI* and LRFI-GWAS, *LRFI*) were performed. Traits with more than 3 QTL differences between the HRFI-GWAS and LRFI-GWAS analyses are highlighted in grey

DFI: daily feed intake*; ADG:* average daily gain*; FCR:* feed conversion ratio*; RFI:* residual feed intake*; carcBFT:* backfat thickness measured on carcass*; a*_GM:* a* measured on the *gluteus medius* muscle*; a*_GS:* a* measured on the *gluteus superficialis* muscle*; b*_GM:* b* measured on the *gluteus medius* muscle*; b*_GS:* b* measured on the *gluteus superficialis* muscle*; L*_GM:* L* measured on the *gluteus medius* muscle*; L*_GS:* L* measured on the *gluteus superficialis* muscle*; pH24h_AD:* pH 24h after slaughter measured on the adductor femoris muscle*; pH24h_GS:* pH 24h after slaughter measured on the *gluteus superficialis* muscle*; pH24h_LM:* pH 24h after slaughter measured on the *longissimus dorsi* muscle*; pH24h_SM:* pH 24h after slaughter measured on the *semimembranosus* muscle*; WHC:* water holding capacity of the *gluteus superficialis* muscle*; MQI:* meat quality index*; LMCcalc:* lean meat content of the carcass*; DP:* carcass dressing percentage*; Belly_W:* belly weight*; BF_W:* backfat weight*; Ham_W:* ham weight*; Loin_W:* loin weight*; Shoulder_W:* shoulder weight

To assess whether the identified QTL regions were identical and shared in the two lines, complementary GWAS analyses were performed per line, using either the set of individuals from G1 to G9 of the HRFI line or the set of individuals from G1 to G9 of the LRFI line. For analyses performed by line, the number of regions detected for a trait could differ between lines. For instance, more loci were detected in the HRFI line for b*_GS, L*_GM and WHC, whilst more regions were detected in the LRFI line for RFI, carcBFT, DP, pH24h_LM and pH24h_AD. In the HRFI line, 46 QTL were identified in 37 regions, and in the LRFI line, 68 QTL were identified in 61 regions. Only 3 regions overlapped in the two lines: on SSC6, a region located between 7 to 10 Mb affected pH24h_LM in LRFI and L*_GS, b*_GS, and MQI in HRFI, on SSC7, a region from 107 to 109 Mb affected L*_GS in HRFI and pH24h_AD in LRFI, and on SSC12, a region located between 7 to 9 Mb affected ADG in HRFI and RFI in LRFI. The two first regions affected highly correlated traits related to meat quality, but the last region affected uncorrelated traits.

Fifteen regions were shared between the 54 regions identified in the Global-GWAS and the 95 unique regions from the analyses per line, with only 3 regions common to the Global-GWAS and HRFI-GWAS analyses, 10 common to Global-GWAS and LRFI-GWAS, and the SSC6 and SSC7 regions described

above detected in the three analyses (Figure 4a). Among these regions only 9 QTL (trait x region) were identified jointly in the Global-GWAS and in one of the Lines-GWAS (Figure 4b), and none was shared in the three analyses. Very few QTL were thus common to the three GWAS (Figure 3). To assess whether a SNP-QTL significant in one analysis reached significance or suggestive thresholds in the other analyses, their *p*-values were compared. First, comparing the Lines-GWAS (Figure 5a), SNP-QTL detected *via* HRFI-GWAS had -$\log_{10}$(*p*-values) generally lower than 1 in the LRFI-GWAS, and none reached the suggestive threshold of 3. Similar results were obtained comparing SNP-QTL of the LRFI-GWAS to their *p*-values with the HRFI-GWAS. For the SNP-QTL significant in the Global-GWAS, the -$\log_{10}$(*p*-values) with the Lines-GWAS were intermediate and exceeded the suggestive threshold for several QTL. It should be noted that for these SNP-QTL, when the -$\log_{10}$(*p*-values) was suggestive in one line, it was lower in the other line.

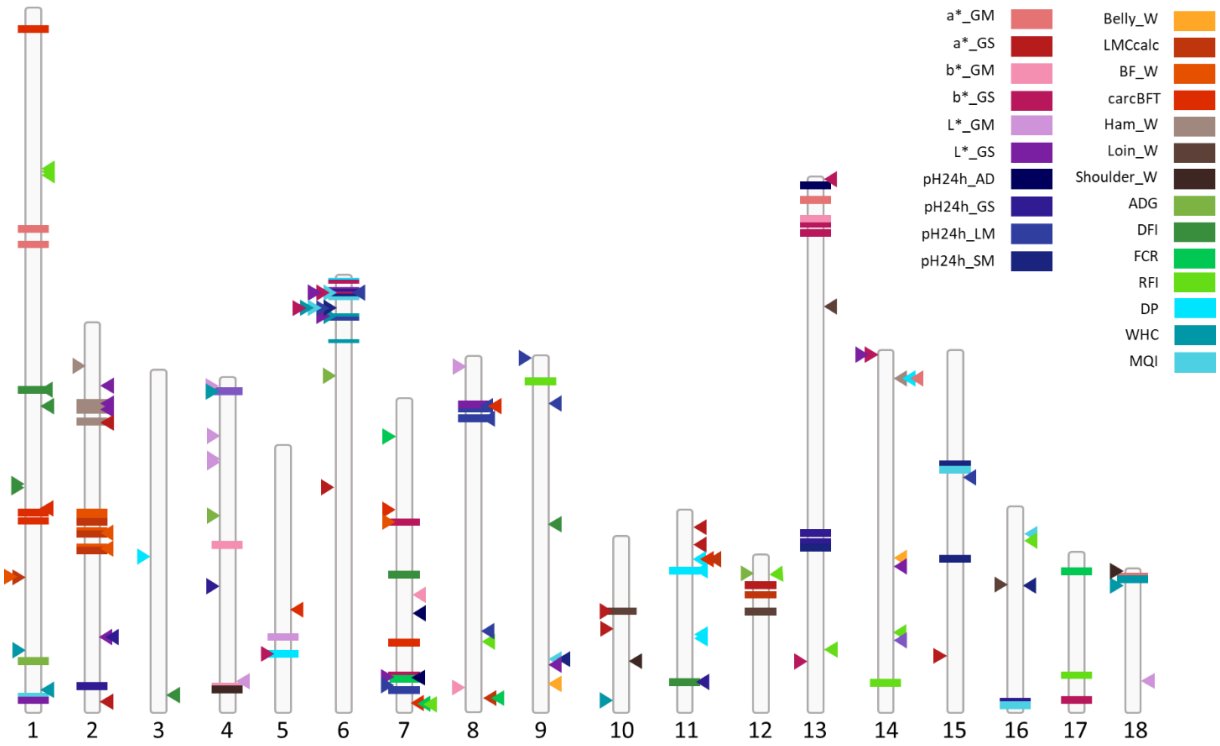

Figure 4 Comparison of GWAS results obtained from Global-GWAS (Global), HRFI-GWAS (HRFI) and LRFI-GWAS (LRFI). (a) Comparison of the number of identical regions and (b) Comparison of the number of identical QTL (trait x region)

In addition, for the SNP-QTL corresponding to the QTL detected in the line analyses (HRFI-GWAS and LRFI-GWAS), the -$\log_{10}$(*p*-values) obtained in the Global-GWAS were also low (Figure 5b), with three-quarters (74.5%) of the SNP-QTL having -$\log_{10}$(*p*-values) lower than 3.

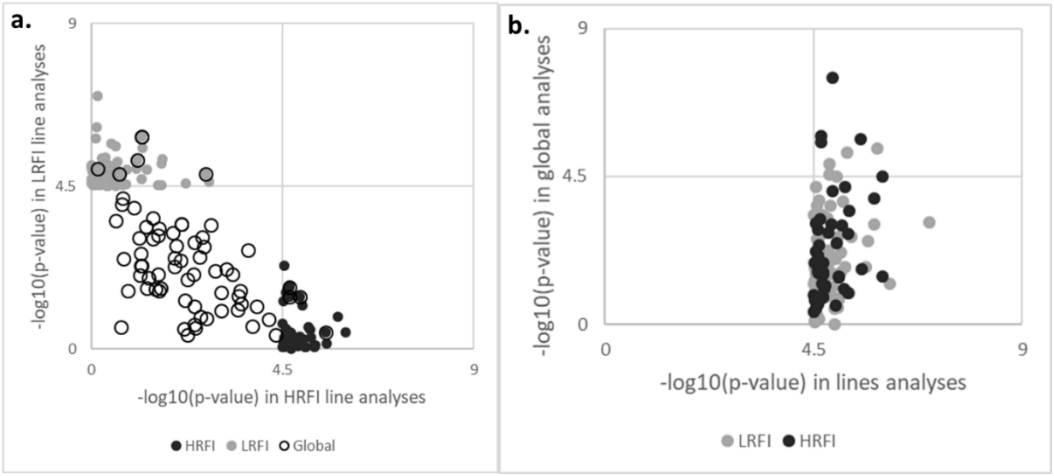

Figure 5 Plot of the -log$_{10}$(*p*-value) of the SNP-QTL. The -log$_{10}$(*p*-value) are obtained in first case with the two lines analyses for all SNP-QTL detected for the lines or the global analyses (a), and in second case obtained with the global analysis for SNP-QTL detected with the GWAS performed per line (b).

Change of allele frequencies over the generations

The allele frequencies of the SNP-QTL detected either in Global-GWAS or in Lines-GWAS were evaluated in G1 to G9 to reflect the informativeness of these GWAS (called G9 hereafter) and in G1. When the SNP-QTL was detected in the Global-GWAS, all response animals were used to compute the frequencies; for SNP-QTL from the Lines-GWAS only the animals of the significant analysis (HRFI-GWAS or LRFI-GWAS) were used. The resulting frequency histograms are shown in Figure 6. With G1 only, 87% of the SNP-QTL of the Global-GWAS had an allele frequency between 0.2 and 0.5, with half of them between 0.4 and 0.5. In addition, 28% of SNP-QTL of the Lines-GWAS have a frequency <0.2 in G1 (Figure 6a), so the distribution of the allele frequencies of the SNP-QTL between low (<0.2) and medium was significantly different between the types of analyses ($P < 0.02$ for a Chi^2 with 1 df). This difference in the distribution of the SNP-QTL allele frequencies between the two types of analyses was preserved in G9 (Figure 6b, $P < 0.005$): 16% of the SNP-QTL of the Global-GWAS had a frequency <0.2, compared to 35% for the SNP-QTL of the Lines-GWAS. In addition, 9% of the SNP-QTL of the Lines-GWAS had a frequency >0.6, no marker reached this frequency among SNP-QTL of the Global-GWAS.

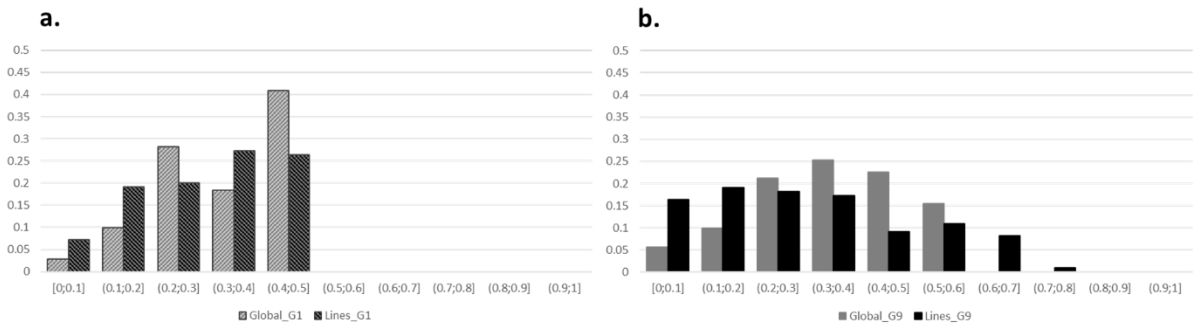

Figure 6 Distribution of SNP-QTL allele frequencies of Global-GWAS (in grey) and Lines-GWAS (in black). Distribution representing individuals from the line of the significant analysis (a) in G1 generation (G1 individuals only) and (b) in G9 generation (G1 to G9 individuals).

In addition to the estimation of the global allelic frequencies, we controlled if in each line the detected SNP-QTL evolved differently according to the type of analysis. First, the differences of allele frequency differences between the HRFI and LRFI lines were estimated in the G1 generation (at the beginning of the selection) (Figure 7). Regardless the analysis in which the SNP-QTL was detected, more than 65% of the SNP-QTL had low line frequency differences (<0.1) and less than 10% of the SNP-QTL had a line frequency difference >0.2. These SNP-QTL were not particularly found in one or the other type of analysis, in both types of analyses.

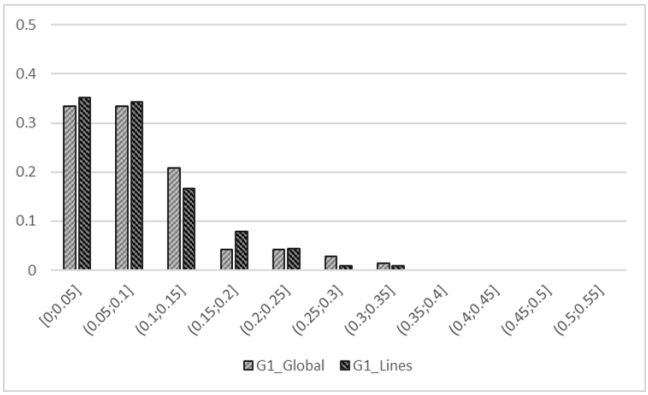

Figure 7 Distribution of allele frequency differences between the lines. The allele frequency differences are the absolute values between lines for SNP-QTL resulting from the Global-GWAS and Lines-GWAS in G1.

To better describe the changes of allele frequency over the generations, frequencies of SNP-QTL from Global-GWAS and Lines-GWAS were then successively estimated in each line by adding data from the next generation to the previous generations: G1 allele frequencies were obtained from G1 individuals alone, G2 allele frequencies were obtained from G1 and G2 individuals etc. Using the 9 resulting frequencies computed in each line, a linear regression of the generation number on the allele frequencies was applied within line. The comparison between lines of the regression coefficients of the allelic frequencies highlighted 4 distinct cases (Figure 8): (1) markers whose frequencies did not change with line selection (slope did not differ from zero Wald test, 5.9%), (2) markers co-selected in the two lines (slopes differed from zero and had identical sign: 22.6%), (3) markers selected in opposite directions in the lines (slopes differed from zero with different signs: 40.3%), and (4) markers whose frequencies changed only in one line (slope different from zero in one line only, 16.7% in LRFI, 14.5% in HRFI). Again no difference in the mean allele frequency evolution was observed for SNP-QTL detected in one or the other type of analysis whatever the situation (*p*-value=0.87 (No-evolution), 0.73

(Co-evolution), 0.50 (Opposite-evolution and 0.70 (One_Line_evolution) for Student T test on the values of the slopes).

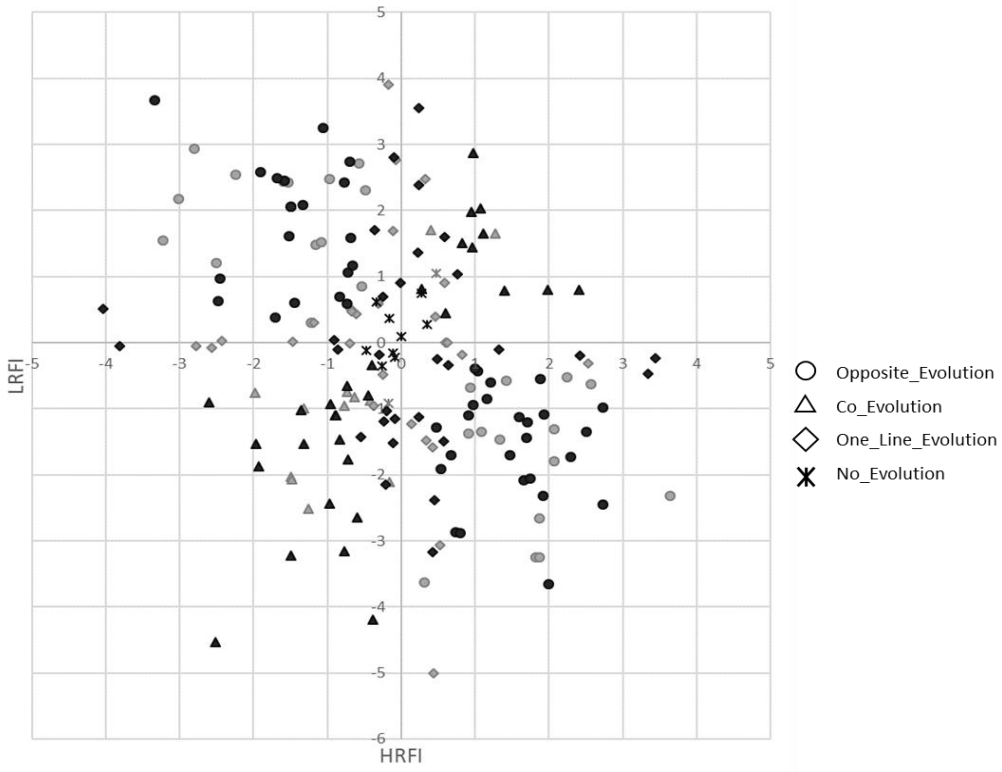

Figure 8 Slopes of the linear regression equations of the allele frequencies on the 9 generations. Slopes were calculated in each line, for all SNP-QTL identified with Global-GWAS (in grey) and Lines-GWAS (in black). Four situations (differentiated by different labels) were identified according to the significance of the slope (different from zero with $p < 0.05$ with a Wald test) in one or the two lines.

For RFI in the two lines, 9 out of the 12 detected QTL corresponded to regions identified with strong line frequencies differences: 3 RFI SNP-QTL showed differences in allelic frequency between lines higher than 0.2 in G1. The other 6 RFI SNP-QTL showed large changes of allelic frequency (regression slope > 0.024/generation). To summarize the changes of SNP-QTL allele frequencies for each trait, an average evolution score between G1 and G9 was computed using the estimated evolution scores of the different SNP-QTL detected for each trait. These averages were between 0.11 (LMCcalc) and 0.24 (RFI). A correlation coefficient of 0.66 was then estimated between the genetic line differences in G9 computed previously for the 24 different traits [27] and these averages (Figure 9).

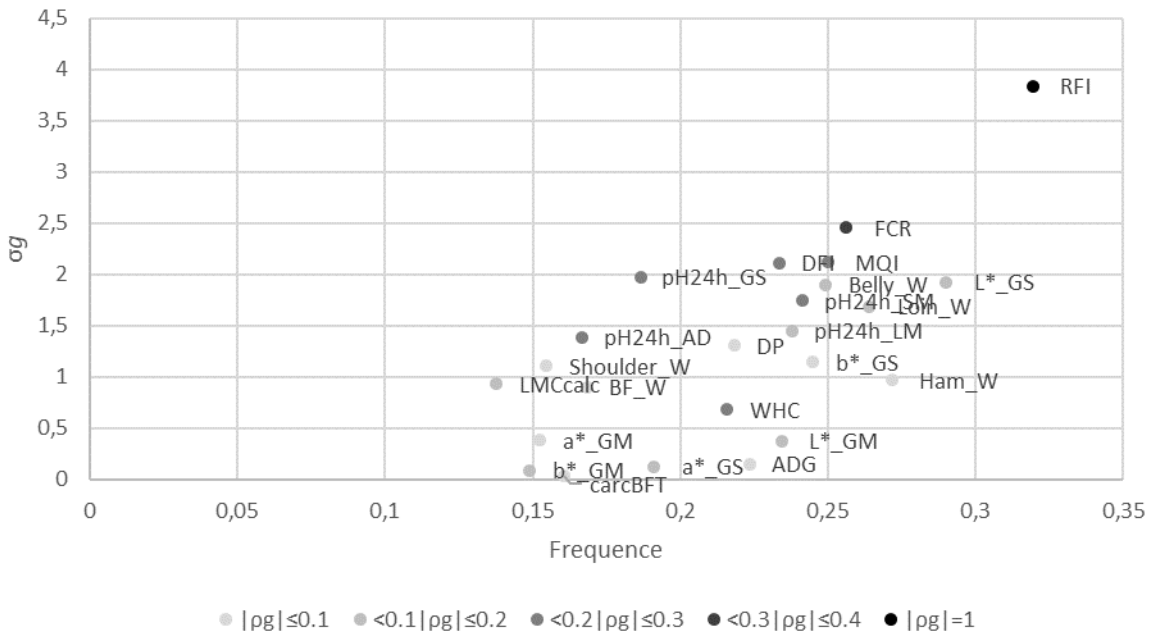

Figure 9 Genetic differences in G9 between the two lines. The genetic differences were expressed in genetic standard deviation of the trait (σ_g) as a function of the average evolution of allelic frequencies in the QTL regions of the trait between the two lines. The magnitude of the genetic correlation between each trait and RFI is indicated with a grey gradient.
DFI: daily feed intake*; ADG:* average daily gain*; FCR:* feed conversion ratio*; RFI:* residual feed intake*; carcBFT:* backfat thickness measured on carcass*; a*_GM:* a* measured on the *gluteus medius* muscle*; a*_GS:* a* measured on the *gluteus superficialis* muscle*; b*_GM:* b* measured on the *gluteus medius* muscle*; b*_GS:* b* measured on the *gluteus superficialis* muscle*; L*_GM:* L* measured on the *gluteus medius* muscle*; L*_GS:* L* measured on the *gluteus superficialis* muscle*; pH24h_AD:* pH 24h after slaughter measured on the adductor femoris muscle*; pH24h_GS:* pH 24h after slaughter measured on the *gluteus superficialis* muscle*; pH24h_LM:* pH 24h after slaughter measured on the *longissimus dorsi* muscle*; pH24h_SM:* pH 24h after slaughter measured on the *semimembranosus* muscle*; WHC:* water holding capacity of the *gluteus superficialis* muscle*; MQI:* meat quality index*; LMCcalc:* lean meat content of the carcass*; DP:* carcass dressing percentage*; Belly_W:* belly weight*; BF_W:* backfat weight*; Ham_W:* ham weight*; Loin_W:* loin weight*; Shoulder_W:* shoulder weight

Discussion

The objective of this study was to identify QTL affecting RFI and production traits in pig lines divergently selected for RFI and to understand if the traits had different genetic backgrounds between the lines. By optimizing the genotyping to reach a good power to detect QTL in the full design and in the two lines separately, QTL were detected for all traits and hypotheses about the trait genetic background in the two lines can be formulated.

Using average parental genotypes to detect QTL

While the use of SNPs chips now enables the genotyping of an individual at a reasonable cost, the genotyping of a design comprising several thousand individuals represents nevertheless a significant

investment. In each generation of our design, at least two parities were produced, one aiming at selecting future breeders, and one to control the responses to the selection on feed consumption, growth and meat quality traits via measurements at the slaughterhouse. After 9 generations of selection, around 2,500 "response animals" had phenotypes. These individuals have the advantage of having individual records for unmeasured traits in breeders (post-mortem measurements). To optimize the costs, we genotyped all 1,632 breeders with MD SNPs chips to exhaustively survey the segregating alleles in the design. In addition, the 32 main contributors to the design were chosen from the G0 sires and dams to be genotyped using the HD SNPs chip, and an imputation step was carried out to have HD genotypes for all breeding individuals. The strong pedigree relationships in the design enabled a very good quality of HD imputation, as they help to better detect long haplotypes used to infer missing SNPs [28]. A second step was carried out, so that each response non-genotyped animal could have a genotype. These non-genotyped animal imputation have been used in cattle [29] as part of genomic evaluations to increase the size of the reference populations. In cattle, the most common situation is to determine by imputation the genotypes of dams of bulls, knowing the genotypes of the maternal grandsire, one (or more) offspring and the sires with which they were mated [30]. In such cases, the strategy takes advantage of the family information (Mendelian rule of allele transmission) and combined with allele frequencies and LD between markers at the population level. In our case, at each generation n, all response animals had both parents genotyped at generation n-1. Given these trio structures, an expected genotype at each position could be deduced from the genotypes of the parents using simple segregation rules: since the genotypes were coded as an allelic dosage for one reference allele, the genotype expectation for each offspring was simply the average of the genotypes of its two parents. As a result, 2,426 animals with genotypes (predicted) and phenotypes were available for subsequent GWAS analyses.

Understanding the differences of detected regions between analyses

The regions detected with each type of analysis (Global- or Lines-GWAS) were very different and only 9 QTL out of 177 were shared between Global-GWAS and Lines-GWAS. The SNP-QTL detected with the Global-GWAS were far from reaching the threshold of significance in the Lines-GWAS. Similarly, most SNP-QTL detected with the Lines-GWAS were far from reaching the threshold of significance in the Global-GWAS. Although the number of individuals included in the Global-GWAS was twice higher than in the line analyses, the addition of individuals belonging to the other line seemed to have reduced the power of detection of QTL segregating in the first line. The SNP-QTL detected in the Global-GWAS or Lines-GWAS differed for their allelic frequencies in G1. This difference remained at the whole line level (G9): more SNPs with low allele frequencies were identified with the Lines-GWAS. The pedigree kinship

matrix was used in the GWAS model to correct for the strong genomic structure of the population. If successful to control the type-I error of the analyses, this classical approach also limits the power of detection of QTL in highly differentiated regions between lines, as their link with the trait variability would be absorbed into the additive genetic component of the model. The Global-GWAS thus essentially allow the detection of regions segregating at intermediate frequencies in both lines. As an alternative, the analyses carried out by line allow detecting regions that got close to fixation with selection in one of the lines. From these results, it seems that the power of detection related to allele frequencies in each line is the main difference between QTL-SNPs detected with the Lines-GWAS and Global-GWAS. Given the power of the design, it is thus likely that the biological pathways involved in RFI variability in the two lines are similar, but with different contributions to the trait in each line, contrary to some previous hypotheses [10, 27].

Comparison with published regions

Among the 12 QTL detected for RFI, three regions are detected close to RFI QTL already published. The region on SSC14 at 130-131 Mb is close to the region described by Duy N. Do et al. [31] who proposed G-protein-coupled receptor kinase 5 (*GRK5*) (129,114,449-129,343,412) as a candidate gene. Wang & al. [32] reported that a GRK5 deficiency led to insulin resistance and hepatic steatosis, and to decreases diet-induced obesity and adipogenesis in mice. In position 131,181,710-131,579,703 *FGFR2* (fibroblast growth factor receptor 2) could also be an interesting candidate. All four FGF receptors and several FGF ligands are present in the intestine and are key players in controlling cell proliferation, differentiation, epithelial cell restitution, and stem cell maintenance. *FGFR2* is expressed in the human ileum and throughout adult mouse intestine [33]. The second region closest to published RFI QTL is the 184-486 Mb interval on SSC13 near QTL reported by Bai et al. [34] and Duy N. Do et al. [31]. In this region *TMPRSS15* (transmembrane serine protease 15) is an interesting candidate gene. This gene encodes an intestinal enzyme responsible for initiating activation of pancreatic proteolytic proenzymes. It catalyzes the conversion of trypsinogen to trypsin, which in turn activates other proenzymes including chymotrypsinogen procarboxypeptidases and proelastases. *TMPRSS15* has been associated to Enterokinase Deficiency, a life-threatening intestinal malabsorption disorder characterized by diarrhea and failure to thrive [35]. On SSC17 two RFI QTL have been published by Duy N. Do et al. [31] close to the *SOGA1* gene (suppressor of glucose, autophagy-associated protein 1, 40,020,107-40,098,992) and by Onteru et al. [10] close to the *DOK5* gene (docking protein 5, 55,391,074-55,541,561). These two QTL surround the region we detected and could correspond to one unique QTL. In position 48,090,077-48,100,816, and in position 48,132,911-48,149,732, respectively, *PLTP* and *ZNF335* genes are additional candidate genes. In human, Coleman et al. [36] identified the

region encoding *ZNF335* as a susceptibility locus for the coeliac disease, a chronic immune-mediated disease triggered by the ingestion of gluten [36]. The PLTP (phospholipid transfer protein) transfers phospholipids from triglyceride-rich lipoproteins to high density lipoprotein (HDL). In addition to regulating the size of HDL particles, this protein may be involved in the cholesterol metabolism. PLTP KO mice absorb less cholesterol than WT mice, and have also deficient secretion by the intestine [37].

Potential pleiotropic effects

The large number of traits recorded in our design and the known genetic correlations between these traits [27] enable the detection of pleiotropic regions, i.e. regions affecting multiple traits. Among the five regions detected for FCR, only the QTL located between 117 Mb and 119 Mb on SSC7 co-localized with a RFI QTL. For the other traits correlated to RFI (DFI, MQI, WHC, pH24h_AD, pH24h_GS, and pH24h_SM traits), only 3 QTL were detected within 10 Mb of the RFI QTL: a QTL at 2 Mb influencing MQI on SSC16 between 11 and 12 Mb, and two QTL on pH24h_AD at 7 Mb and 10 Mb of QTL for RFI located at 113-114 Mb on SSC14 and 107-109 Mb on SSC7, respectively. Compared to the previously published QTL regions for RFI, we identified a QTL influencing FCR in a region described by Onteru et al. [10] between 15 and 16 Mb on SSC7, a QTL for pH24h_SM in the 80 and 81Mb interval on SSC15 described by Duy N Do et al. [31], and a QTL for DFI in the region described by Y M Guo et al. [38] on SSC3 between positions 126 and 128Mb. Despite the reported correlations between these traits and RFI, among the 52 QTL detected in our study for DFI, MQI, WHC, pH24h_AD, pH24h_GS, and pH24h_SM, only seven co-located with RFI QTL identified in our study or in previously published studies.

Changes of QTL allele frequencies and trait responses to selection

The allele frequencies of the majority of the detected regions changed between the G1 and G9 generations, with more than 70% of the regions for which SNP-QTL evolved in opposite directions or in a one line only. However, the magnitude of allelic changes of the QTL regions varied from one trait to the next, and was strongly correlated with line differences previously reported in G9 [27]. Indeed, the regions with the highest allele frequency changes were detected for RFI, which was trait used for selection. For the other traits, the higher the genetic correlation with RFI, the higher the frequency variation of the associated QTL regions. As a result, QTL affecting FCR, DFI and MQI had the highest frequency changes with generations. The responses of QTL affecting meat quality traits are consistent with the high and early responses to selection previously detected in this experimental population for these traits [5]. Altogether, our analyses underline a clear relationship between the quantitative responses to selection of the traits and changes of alleles frequencies in some QTL regions, certainly

pointing out chromosomic regions that were selected during the experiment, whereas in such populations of low effective size and strong directional selection, detecting selection signatures with standard methodologies [39] can have low power due to the major effect of drift on the changes of the allele frequencies. However, recently developed new methods, based on genetic time series could provide new insights for the detection of regions under selection in small populations [40].

Conclusion

This study aimed at characterizing the molecular architecture of RFI in two lines divergently selected for this trait. Besides efficiently detecting known and new QTL regions, the combination of GWAS carried out per line or simultaneously using all individuals allowed the identification of candidate regions of the genome under selection, which can explain the responses to selection of different traits reported before. Analyzing the allelic frequencies from G1 to G9, we concluded that the majority of the QTL regions responded to selection in a divergent way in the lines, and that the same metabolic pathways were certainly involved in both lines. Several new regions determining RFI variability were identified in this study and new candidate genes were proposed to complement the data acquired in previous published analyses.

Additionnal file 1

Animals quality control	60K	70K	650K	MD/HD imputation	HD predicted
Total animals	286	711	32	-	-
CR* animals deleted	0	10	0	-	-
Animals used	286	701	32	987	2.426

Number of animals used for the analyses after quality control. Details of the number of aniamls before and after application of filter on the call rate (CR) were given for chips (60K, 70K and 650K SNPs chips), imputation levels (MD/HD imputation) and average genotypes calculated from the genotypes of both parents (HD predicted).

Additionnal file 2

SNP quality control	60K	70K	650K	MD imputation	HD imputation	HD predicted
Total SNPs	64.232	68.516	658.692	66.988	570.447	570.447
CF* SNPs deleted	5.776	5.323	53.735	-	-	-
MAF* SNPs deleted	9.125	6.568	45.852	-	-	-
Total SNPs deleted	15.114	11.891	99.587	-	-	-
SNPs after filtering	49.118	56.625	559.105	-	-	-

Number of SNPs used for the analyses after quality control. Details of the number of SNPs before and after application of filters on the call frequency (CF) and the frequency of minor allele (MAF) were given for chips (60K, 70K and 650K SNPs chips), imputation levels (MD imputation and HD imputation) and average genotypes calculated from the genotypes of both parents (HD predicted).

Additionnal file 3

Cf. Annexe 1.

QTL regions detected with the three groups of association studies. These QTL regions were found from the full population (Global-GWAS) and from each line separately (HRFI-GWAS and LRFI-GWAS)
DFI: daily feed intake*; ADG:* average daily gain*; FCR:* feed conversion ratio*; RFI:* residual feed intake*; carcBFT:* backfat thickness measured on carcass*; a*_GM:* a* measured on the *gluteus medius* muscle*; a*_GS:* a* measured on the *gluteus superficialis* muscle*; b*_GM:* b* measured on the *gluteus medius* muscle*; b*_GS:* b* measured on the *gluteus superficialis* muscle*; L*_GM:* L* measured on the *gluteus medius* muscle*; L*_GS:* L* measured on the *gluteus superficialis* muscle*; pH24h_AD:* pH 24h after slaughter measured on the adductor femoris muscle*; pH24h_GS:* pH 24h after slaughter measured on the *gluteus superficialis* muscle*; pH24h_LM:* pH 24h after slaughter measured on the *longissimus dorsi* muscle*; pH24h_SM:* pH 24h after slaughter measured on the *semimembranosus* muscle*; WHC:* water holding capacity of the *gluteus superficialis* muscle*; MQI:* meat quality index*; LMCcalc:* lean meat content of the carcass*; DP:* carcass dressing percentage*; Belly_W:* belly weight*; BF_W:* backfat weight*; Ham_W:* ham weight*; Loin_W:* loin weight*; Shoulder_W:* shoulder weight

Declarations

Ethics approval and consent to participate

All pigs were reared in compliance with national regulations and according to procedures approved by the French Veterinary Services at INRA experimental facilities. The care and use of pigs were performed following the guidelines edited by the French Ministries of High Education, Research and Innovation, and of Agriculture and Food (http://ethique.ipbs.fr/sdv/charteexpeanimale.pdf).

Consent for publication

Not applicable

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This study and the two first authors were financially supported by the French National Research Agency via the PIG_FEED and MicroFeed projects, under grants ANR-08-GENM-038 and ANR-16-CE20-0003.

Authors' contributions

ED performed the statistical analyses and wrote the first draft of the paper. YB and KF organized the data acquisition. ED and YL performed the imputation and quality control of the genotypic data. ED, AA, YL, HG and JR participated in the design of the study. JR and HG provided scientific supervision. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank (*i*) the experimental farm staff for data collection, samples management and breeding of the animals and (*ii*) both technology platforms, CRCT and Gentyane, for the genotyping.

References

1. McGlone J, Pond WG. Pig Production: Biological Principles and Applications. Cengage Learning; 2003.

2. Soleimani T, Gilbert H. Evaluating environmental impacts of selection for residual feed intake in pigs. animal. 2020;:1–11.

3. Webb AJ, King JWB. Selection for improved food conversion ratio on ad libitum group feeding in pigs. Animal Science. 1983;37:375–85.

4. Koch RM, Swiger LA, Chambers D, Gregory KE. Efficiency of Feed Use in Beef Cattle. J Anim Sci. 1963;22:486–94.

5. Gilbert H, Bidanel J-P, Gruand J, Caritez J-C, Billon Y, Guillouet P, et al. Genetic parameters for residual feed intake in growing pigs, with emphasis on genetic relationships with carcass and meat quality traits. J Anim Sci. 2007;85:3182–8.

6. Cai W, Casey DS, Dekkers JCM. Selection response and genetic parameters for residual feed intake in Yorkshire swine. J Anim Sci. 2008;86:287–98.

7. Drouilhet L, Achard CS, Zemb O, Molette C, Gidenne T, Larzul C, et al. Direct and correlated responses to selection in two lines of rabbits selected for feed efficiency under ad libitum and restricted feeding: I. Production traits and gut microbiota characteristics. J Anim Sci. 2016;94:38–48.

8. 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. Scientific Reports. 2018;8:558.

9. 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.

10. Onteru SK, Gorbach DM, Young JM, Garrick DJ, Dekkers JCM, Rothschild MF. Whole Genome Association Studies of Residual Feed Intake and Related Traits in the Pig. PLOS ONE. 2013;8:e61756.

11. Ding R, Yang M, Wang X, Quan J, Zhuang Z, Zhou S, et al. Genetic Architecture of Feeding Behavior and Feed Efficiency in a Duroc Pig Population. Front Genet. 2018;9. doi:10.3389/fgene.2018.00220.

12. 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 Res. 2019;47:D701–10.

13. Sosa-Madrid BS, Santacreu MA, Blasco A, Fontanesi L, Pena RN, Ibáñez-Escriche N. A genomewide association study in divergently selected lines in rabbits reveals novel genomic regions associated with litter size traits. Journal of Animal Breeding and Genetics. 2020;137:123–38.

14. Daumas G. Taux de muscle des pièces et appréciation de la composition corporelle des carcasses. :7.

15. Charpentier J, Monin G, Ollivier L. Correlations between carcass characteristics and meat quality in Large White pigs. Proceedings of the 2nd International Symposium on Conditions and Meat Quality of Pigs. 1971. https://agris.fao.org/agris-search/search.do?recordID=US201302309831. Accessed 5 Oct 2020.

16. Tribout T, Caritez J-C, Gogué J, Gruand J, Bouffaud M, Billon Y, et al. Estimation, par utilisation de semence congelée, du progrès génétique réalisé en France entre 1977 et 1998 dans la race porcine Large White : résultats pour quelques caractères de production et de qualité des tissus gras et maigres. :8.

17. Noblet J, Karege C, Dubois S, van Milgen J. Metabolic utilization of energy and maintenance requirements in growing pigs: effects of sex and genotype. J Anim Sci. 1999;77:1208–16.

18. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics. 2007;81:559–75.

19. Warr A, Affara N, Aken B, Beiki H, Bickhart DM, Billis K, et al. An improved pig reference genome sequence to enable pig genetics and genomics research. Gigascience. 2020;9. doi:10.1093/gigascience/giaa051.

20. Sargolzaei M, Chesnais JP, Schenkel FS. A new approach for efficient genotype imputation using information from relatives. BMC Genomics. 2014;15:478.

21. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.

22. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012;44:821–4.

23. Aliakbari A, Delpuech E, Labrune Y, Riquet J, Gilbert H. The impact of training on data from genetically-related lines on the accuracy of genomic predictions for feed efficiency traits in pigs. Genet Sel Evol. 2020;52:57.

24. Zhou X, Carbonetto P, Stephens M. Polygenic Modeling with Bayesian Sparse Linear Mixed Models. PLOS Genetics. 2013;9:e1003264.

25. Gao X. Multiple testing corrections for imputed SNPs. Genetic Epidemiology. 2011;35:154–8.

26. Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: a review. Plant Methods. 2013;9:29.

27. Gilbert H, Billon Y, Brossard L, Faure J, Gatellier P, Gondret F, et al. Review: divergent selection for residual feed intake in the growing pig. animal. 2017;11:1427–39.

28. Ullah E, Mall R, Abbas MM, Kunji K, Nato AQ, Bensmail H, et al. Comparison and assessment of family- and population-based genotype imputation methods in large pedigrees. Genome Res. 2019;29:125–34.

29. Bouwman AC, Hickey JM, Calus MP, Veerkamp RF. Imputation of non-genotyped individuals based on genotyped relatives: assessing the imputation accuracy of a real case scenario in dairy cattle. Genet Sel Evol. 2014;46:6.

30. Pimentel EC, Wensch-Dorendorf M, König S, Swalve HH. Enlarging a training set for genomic selection by imputation of un-genotyped animals in populations of varying genetic architecture. Genetics Selection Evolution. 2013;45:12.

31. Do DN, Ostersen T, Strathe AB, Mark T, Jensen J, Kadarmideen HN. Genome-wide association and systems genetic analyses of residual feed intake, daily feed consumption, backfat and weight gain in pigs. BMC Genetics. 2014;15:27.

32. Wang L, Shen M, Wang F, Ma L. GRK5 ablation contributes to insulin resistance. Biochemical and Biophysical Research Communications. 2012;429:99–104.

33. Danopoulos S, Schlieve CR, Grikscheit TC, Alam DA. Fibroblast Growth Factors in the Gastrointestinal Tract: Twists and Turns. Developmental Dynamics. 2017;246:344–52.

34. Bai C, Pan Y, Wang D, Cai F, Yan S, Zhao Z, et al. Genome-wide association analysis of residual feed intake in Junmu No. 1 White pigs. Animal Genetics. 2017;48:686–90.

35. Holzinger A, Maier E, Bück C, Mayerhofer P, Kappler M, Haworth J, et al. Mutations in the Proenteropeptidase gene are the molecular cause of congenital enteropeptidase deficiency. 2002. doi:10.1086/338456.

36. Coleman C, Quinn EM, Ryan AW, Conroy J, Trimble V, Mahmud N, et al. Common polygenic variation in coeliac disease and confirmation of ZNF335 and NIFA as disease susceptibility loci. European Journal of Human Genetics. 2016;24:291–7.

37. Liu Ruijie, Iqbal Jahangir, Yeang Calvin, Wang David Q.-H., Hussain M. Mahmood, Jiang Xian-Cheng. Phospholipid Transfer Protein–Deficient Mice Absorb Less Cholesterol. Arteriosclerosis, Thrombosis, and Vascular Biology. 2007;27:2014–21.

38. Guo YM, Zhang ZY, Ma JW, Ai HS, Ren J, Huang LS. A genomewide association study of feed efficiency and feeding behaviors at two fattening stages in a White Duroc × Erhualian F2 population. J Anim Sci. 2015;93:1481–9.

39. Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B. Using haplotype differentiation among hierarchically structured populations for the detection of selection signatures. Genetics. 2013;193:929–41.

40. Paris C, Servin B, Boitard S. Inference of Selection from Genetic Time Series Using Various Parametric Approximations to the Wright-Fisher Model. G3: Genes, Genomes, Genetics. 2019;9:4073–86.

V. Comparaison des régions QTL identifiées pour la CMJR avec celles publiées pour d'autres espèces de rentes

Le dispositif expérimental de sélection divergente avec lequel nous avons pu réaliser de multiples études GWAS a permis l'identification de 186 régions QTL pour les 24 caractères étudiés. Certaines des régions QTL identifiées avec notre dispositif de sélection divergente ont été reportées pour les mêmes caractères ou pour d'autres caractères dans la base de données pigQTLdb (Hu et al., 2019). Mais des études portant sur l'architecture génétique de la CMJR ont également été réalisées pour d'autres espèces de rente comme les bovins et les poulets. Nous avons donc souhaité réaliser une comparaison des régions QTL pour la CMJR identifiées soit avec notre dispositif soit à partir de dispositifs bovins et poulets. Pour réaliser cette comparaison, j'ai effectué une récupération sur la base de données Cattle QTLdb et Chicken QTLdb de toutes les régions QTL répertoriées pour la CMJR et issues d'analyses GWAS. Afin de comparer la localisation de ces différents QTL, j'ai sélectionné l'ensemble des gènes présents dans les intervalles de localisation des QTL bovins et aviaires annotés sur les génomes de référence de ces espèces (Cow, ARS-UCD1.2 et Chicken, GRCg6a) et j'ai alors recherché leur position sur le génome du porc. Cette approche qui utilise les gènes annotés comme ancrage entre les génomes, m'a ainsi permis d'identifier les régions orthologues porcines de l'ensemble des QTL pour la CMJR bovin et aviaire. La figure 39 résume les positions respectives des QTL détectés dans notre étude pour la CMJR et les positions porcines orthologues des QTL détectés chez les bovins et le poulet. Compte tenu de la faible précision de localisation des régions QTL, 3 régions pourraient être analysées plus précisément : sur le SSC1 le QTL localisé aux alentours de 77 Mb est à proximité d'un QTL détecté en poulet localisé sur le GGA3 à 52-65 Mb (Mignon-Grasteau et al., 2015a) ; sur le SCC12 le QTL détecté à environ 9 Mb est à proximité d'un QTL bovin localisé sur le BTA19 à 59 Mb (Seabury et al., 2017) ; enfin le QTL localisé sur le SSC14 à environ 131 Mb est à proximité d'un QTL bovin localisé sur le BTA26 à 46 Mb (Lu et al., 2018). Cette analyse très préliminaire pourrait être affiner afin (*i*) d'évaluer l'apport de la cartographie comparée pour la caractérisation fine de région QTL et (*ii*) d'évaluer si les processus biologiques sous-jacents à la CMJR font appel à des voies métaboliques identiques ou non selon les espèces.

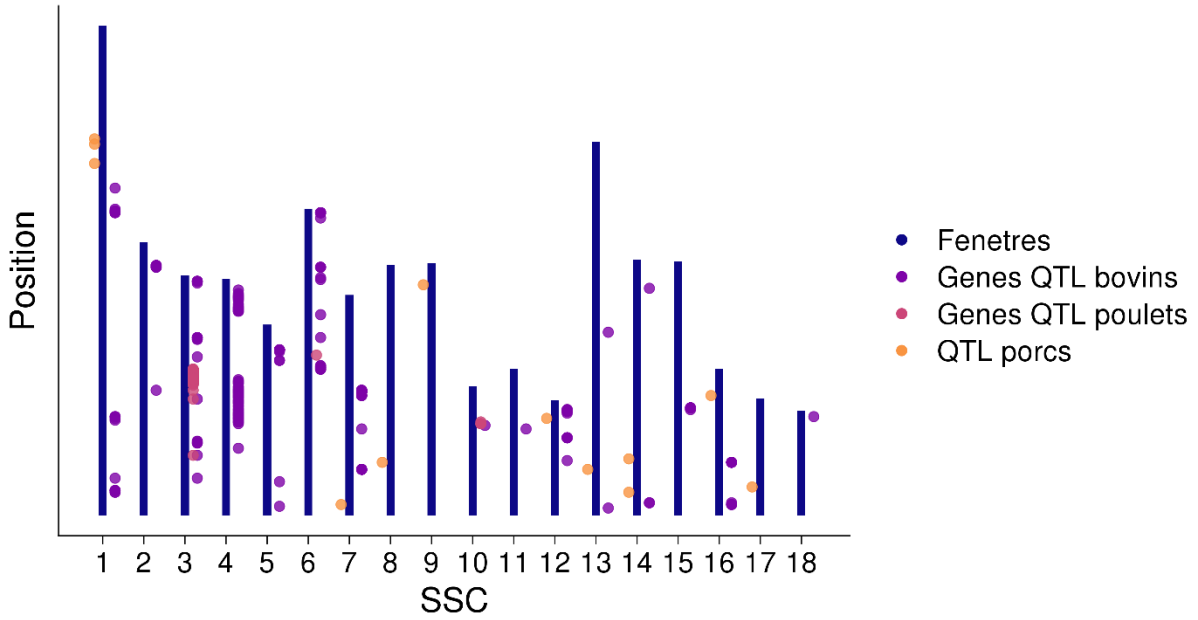

Figure 39 : Caryotype représentant les 18 autosomes et où chaque chromosome est caractérisé par le nombre de fenêtres de 1Mb. La position des gènes présents dans les QTL identifiés en bovin (rond violet) et en poulet (rond rose) ont été reportés sur le côté droit des chromosomes et les régions QTL détectées pour la CMJR dans mon analyse sont représentées sur le côté gauche des chromosomes (rond orange)

VI. Discussion

Au regard des données disponibles, les GWAS ont été effectuées avec des génotypes parentaux moyens. Notre approche est novatrice pour son utilisation dans la recherche de régions en ségrégation via des analyses GWAS. Dans l'article précédemment présenté, l'ensemble des individus de la seconde parité ne possédaient que des génotypes parentaux moyens. Afin de conforter notre approche il serait intéressant de comparer la puissance et la précision de notre dispositif par rapport à l'utilisation de génotypes réels Dans le cadre du projet Micro-Feed, 397 animaux appartenant à ce dispositif ont été génotypés à l'aide de la puce MD 70K. A partir de ce lot d'individus, nous avons calculé une corrélation entre les vrais génotypes et les génotypes parentaux moyens du chromosome 18 et obtenu une corrélation moyenne de 0,8314. Cette valeur est forte mais elle reflète également une potentielle perte en puissance de détection.

A partir de ces données il serait intéressant de réaliser des simulations de détection de QTL. Nous pourrions choisir un SNP, considérer qu'il représente le QTN recherché (en simulant un effet de substitution allèle sur un caractère) et faire des GWAS soit à partir des génotypes réels disponibles soit à partir des génotypes moyens. Plusieurs situations, avec des effets du QTN et des fréquences alléliques du QTN, pourraient être testées. En réitérant ces simulations un millier de fois par situation nous aurions ainsi une première estimation de perte de puissance (nombre de fois que le QTN est détecté avec les vrais génotypes et non détecté avec les génotypes moyens) et d'impact sur la précision (position par rapport au QTN du marqueur le plus significatif pour les GWAS réalisés avec les vrais génotypes ou les génotypes moyens).

Enfin les dernières comparaisons intéressantes à réaliser, seraient de confronter les résultats que nous avons obtenus avec les résultats d'une analyse GWAS qui aurait été faite à partir des vrais génotypes des individus reproducteurs et les valeurs génétiques d'élevage estimées (EBV) de ces individus pour les 24 caractères estimés à partir des phénotypes des individus contrôles (P2). En théorie les résultats devraient être comparables mais restent néanmoins à vérifier.

Chapitre 3 : Intégration de données génétiques, transcriptomiques et d'annotations fonctionnelles

Dans le chapitre précédent, 186 régions QTL ont été identifiées et une recherche dans la littérature sur la fonction des gènes annotés dans ces régions a été réalisée pour identifier des gènes candidats positionnels et fonctionnels. Si cette analyse est une première approche d'intégration de données d'annotation, elle reste uniquement basée sur une analyse gène par gène et région par région à partir de données génériques publiées (non spécifiques du dispositif expérimental). Afin d'affiner ces travaux d'exploitation de données fonctionnelles pour la caractérisation des régions QTL deux voies peuvent être explorées : la première est basée sur l'exploitation de données fonctionnelles issues des animaux du dispositif expérimental, la seconde est l'exploration simultanée de l'ensemble des régions QTL détecté pour un caractère. Dans le cadre de cette thèse, aucune étude transcriptomique n'a été réalisée ni sur les animaux sélectionnés et génotypés (P1) ni sur les animaux avec lesquels les GWAS ont été réalisées (P2) ; néanmoins un lot de 48 individus issus de la 8ème génération de sélection a été étudié par Gondret et al. (Gondret et al., 2017) via une analyse transcriptomique de 4 tissus prélevés sur chacun des individus. Ces données transcriptomiques sont donc directement issues des lignées divergentes sélectionnées sur la CMJR que nous étudions et permettent ainsi d'explorer la première approche basée sur l'analyse des données fonctionnelles issues du même dispositif pour définir plus précisément les gènes candidats des régions QTL. La seconde approche s'appuie essentiellement sur l'analyse des gènes positionnés dans les régions QTL via les connaissances fonctionnelles disponibles dans les bases de données génériques pour l'ensemble des gènes.

L'objectif de ce second chapitre a donc été d'affiner la recherche de gènes candidats dans les régions QTL pour améliorer la caractérisation du phénotype de l'efficacité alimentaire, mais aussi plus largement, d'identifier des processus biologiques sous-jacents aux lignées divergentes via les données d'expression obtenues sur le même dispositif et en réalisant des études d'enrichissement.

I. Introduction

Afin d'affiner notre compréhension de l'architecture génétique du caractère CMJR nous avons choisi d'exploiter et de combiner trois types de données : les résultats obtenus à partir des analyses GWAS, soit 186 régions QTL pour 24 caractères (« GWAS »), les données transcriptomiques provenant des mêmes lignées (« DEG »), dont l'analyse des données a été réalisée par Gondret et al. (Gondret et al., 2017) et les données d'annotations fonctionnelles issues de la base de données Gene Ontologie (GO) regroupant les gènes selon leur appartenance à un même processus biologique (« Annotation »). Les analyses réalisées dans ce chapitre sont illustrées via un diagramme de Venn (Figure 40) : 4 approches ont été menées correspondant aux 4 intersections entre chaque type d'information. La première approche (1) est la combinaison des données transcriptomiques avec l'annotation fonctionnelle des gènes disponibles dans les bases de données. Cette partie correspond donc à l'étude de Gondret et al. (Gondret et al., 2017). La seconde approche (2) correspond à l'analyse combinée des DEG et des régions QTL dans le but d'identifier des DEG aux positions ou à proximité d'une région QTL en faisant l'hypothèse que le QTN affecterait directement la transcription du gène responsable de la variabilité du caractère. La troisième partie (3) a pour but d'identifier des voies métaboliques enrichies en gènes présents dans les régions QTL. L'avantage de cette approche est de pouvoir combiner dans une même recherche la totalité des gènes des régions QTL, a contrario des analyses gène par gène et/ou région par région effectuées dans le chapitre 1. Enfin la dernière approche (4) est destinée à combiner les trois jeux de données et à rechercher des voies métaboliques enrichies en gènes localisés dans des régions génétiques candidates et présentant un différentiel d'expression entre les deux lignées. Chacune de ces 4 combinaisons fait l'objet d'une sous-partie de ce chapitre de thèse.

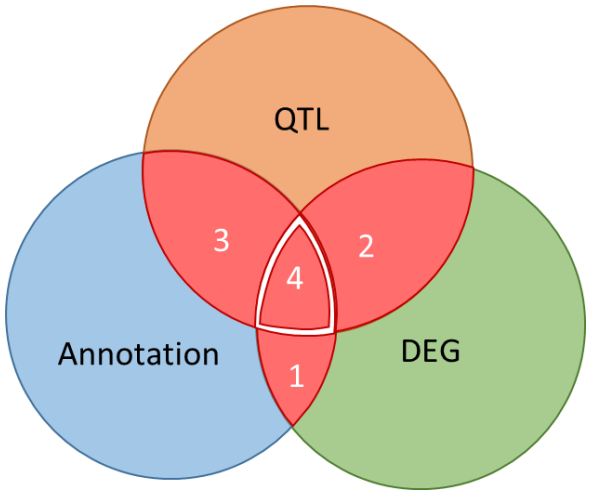

Figure 40 : Diagramme de Venn de la combinaison des 3 types de données (données génétiques : « QTL », données transcriptomiques : « DEG » et les données d'annotation : « Annotation ») pris en compte dans notre étude, afin d'affiner l'identification de gènes candidats pour la cartographie fine des régions QTL en lien avec l'efficacité alimentaire.

II. Analyse d'enrichissement à partir de données transcriptomiques obtenues sur les lignées CMJR, Gondret et al. 2017

Cette première analyse correspond aux travaux réalisés par Gondret et al. (Gondret et al., 2017). L'objectif de ce travail était de caractériser les voies métaboliques associées aux différences d'efficacité alimentaire dans 4 tissus. Bien que déjà évoqué dans le chapitre bibliographique, nous avons estimé que présenter un résumé de cette étude et des résultats obtenus dans cette partie facilitera la lecture des parties suivantes. De plus afin de faciliter la comparaison des résultats de cette sous-partie avec les résultats des sous-parties suivantes j'ai réalisé de nouveau une analyse d'enrichissement à partir de la liste des DEG à l'aide de l'outils GSEA.

1. Le dispositif et les données disponibles

L'analyse de Gondret et al. (Gondret et al., 2017) a été réalisée à partir de 48 individus (24 par lignée) de la génération 8 nés dans une quatrième parité de croisements entre les reproducteurs G7. Ces animaux sont des mâles castrés et correspondent donc à des animaux contemporains des animaux utilisés dans l'étude GWAS du chapitre 1. Pour ces animaux, seules des données transcriptomiques ont été prises en compte dans ce travail de thèse. A 132 jours d'âge les 48 animaux ont été abattus et 4 tissus ont été prélevés, le foie, le muscle de la longe correspondant au muscle à contraction rapide de la partie lombaire, le tissu adipeux sous-cutanés dorsal (SCAT) et le tissu adipeux périrénal (PRAT). A partir de ces tissus, 192 échantillons d'ARN ont été analysés à l'aide de la puce transcriptomique porcine Agilent-037880/INRA Sus scrofa 60K V1 (les résultats sont disponibles dans GEO http://www.ncbi.nlm.nih.gov/geo/ avec le numéro d'accession GPL16524). La puce transcriptomique a été personnalisée via un enrichissement en gènes du système immunitaire. Ces puces transcriptomiques contiennent 61 265 sondes dont 1 319 sont des sondes contrôles (contrôles Agilent, correspondant à des contrôles négatifs ou à des sondes localisées dans les coins des puces et utilisées pour l'alignement des puces) et 60 306 sondes qui correspondent à des oligonucléotides ciblant des transcrits. Les données d'expression brutes obtenues, correspondant aux médianes des intensités des pixels constituant chaque sonde, ont été soumises à un filtre de qualité basé sur 4 critères : l'intensité, l'uniformité, la saturation et la détection des valeurs aberrantes. Enfin, les intensités des sondes filtrées ont été transformées en log2(Cy3) et pour chaque échantillon, une étape de normalisation par centrage de la médiane, c'est-à-dire en soustrayant la valeur médiane de toutes les valeurs brutes a été réalisée. Au total, 12 501 gènes possèdent une valeur d'expression. Afin de définir la liste des DEG entre les lignées un même seuil a été appliqué pour chaque tissu correspondant à une p-value < 0,01 pour un ratio d'expression > 1,1. La liste de DEG comprenant les gènes surexprimés (ratio ≥ 1,1) et sous-exprimés (ratio ≤ 0,91) dans la lignée CMJR- en comparaison à la lignées CMJR+ comprend 2 494

DEG identifiés dans le muscle, 1 101 DEG dans le foie, 913 DEG dans le PRAT et 646 DEG dans le SCAT (Figure 41).

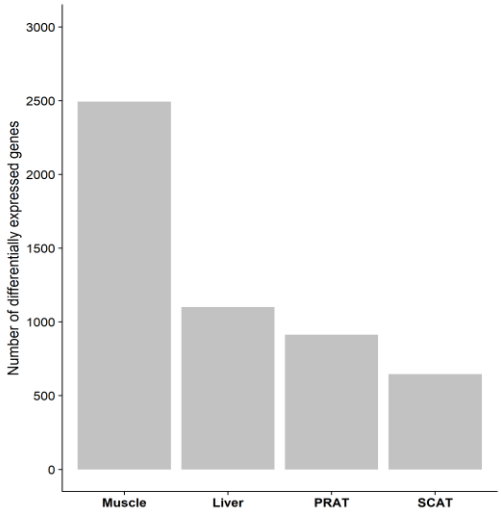

Figure 41 : Barplot du nombre de DEG par tissu qui sont issus de l'analyse des données transcriptomiques présenté dans l'article de Gondret et al. (Gondret et al., 2017) dont le muscle (« Muscle »), le foie (« Liver »), le tissu adipeux périrénal (« PRAT ») et le tissu adipeux sous cutané (« SCAT »).

Si on combine l'ensemble de ces DEG, une liste de 3 684 DEG a ainsi été obtenue dont la majorité possède une expression supérieure dans la lignée CMJR- comparativement à la lignée CMJR+ (Figure 42).

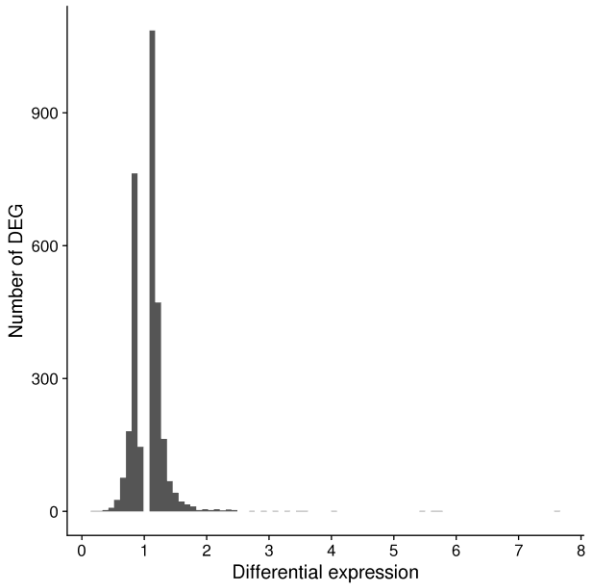

Figure 42 : Distribution du nombre de DEG par rapport à leur différentiel d'expression

Le nombre de gènes différentiellement exprimés dans un seul tissu était néanmoins majoritaire dans les tissus maigres comparé au nombre de DEG détectés dans le gras, représenté par l'analyse de 2 tissus adipeux (Figure 43). La sélection de ces lignées a été menée avec l'objectif de maintenir une adiposité constante dans les deux lignées et cela s'est traduit dans cette étude transcriptomique par une faible différence entre les lignées dans l'expression des gènes du tissu SCAT. A l'inverse, le transcriptome du muscle est celui qui est le plus affecté par la sélection divergente pour la CMJR mais ce résultat est cohérent en raison du rôle majeur du muscle dans l'homéostasie énergétique de chaque individu et de l'utilisation globale de l'énergie. La comparaison des DEG partagés a permis d'identifier 147 gènes communs aux quatre tissus (Figure 43).

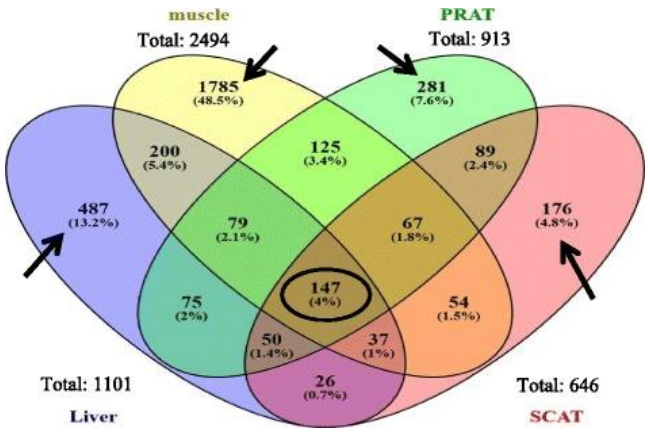

Figure 43 : Diagramme de Venn des DEG identifiés dans les 4 tissus étudiés, le muscle (« Muscle »), le foie (« Liver ») et les 2 tissus adipeux (« PRAT » et « SCAT ») d'après les résultats d'analyse des données transcriptomiques présentés dans l'article de Gondret et al. (Gondret et al., 2017).

2. Analyse d'enrichissement

a. Analyses réalisées par Gondret et al

A partir de la liste de DEG, une recherche de termes GO pour des processus biologiques a été réalisée à l'aide de l'outil DAVID (Database for Annotation, Visualization and Integrated Discovery) (Huang et al., 2009) et des réseaux de régulation candidats ont été visualisés à l'aide de l'outil QIAGEN's Ingenuity®. A partir des 147 DEG communs aux 4 tissus, 8 voies métaboliques ont été mises en évidence. Le transport intracellulaire et la réponse immunitaire sont les processus biologiques les plus enrichis. A cela s'ajoute le processus métabolique des lipides membranaires, le processus catabolique des macromolécules, l'élongation de la traduction protéique, le processus métabolique des phospholipides, la réponse au stress oxydatif et un groupe de voies métaboliques liées à la mort cellulaire. Ces 8 voies sont considérées comme les principales voies métaboliques de réponses biologiques jouant un rôle dans la différence d'efficacité alimentaire. Via une analyse tissu par tissu,

41 clusters fonctionnels (termes GO) ont été identifiés dans le muscle, le foie ou les tissus adipeux. La liste des termes GO est rapportée dans la table 7.

Table 7 : Liste des principaux processus biologiques mis en évidence pour chaque tissu et présentés dans l'article de Gondret et al. (Gondret et al., 2017).

Tissue	GO terms	Nb DEG	E score	*P*-value
	Over-expressed in pigs from the low RFI line			
Muscle	GO:0006412 ~ translation	97	39.8	<0.001
Muscle	GO:0022613 ~ ribonucleoprotein complex biogenesis	33	4.3	<0.001
Muscle	GO:0003012 ~ muscle system contraction process	26	4.1	<0.001
Muscle	GO:0006091 ~ generation of precursor metabolites and energy	31	3.1	<0.001
Muscle	(GO:0006006 ~ glucose metabolic process)	(26)		(0.007)
Muscle	GO:0032268 ~ regulation of cellular protein metabolic process	52	2.8	0.001
Muscle	(GO:0006417 ~ regulation of translation)	(23)		(<0.001)
Muscle	GO:0030163 ~ protein catabolic process	67	2.7	<0.001
Muscle	(GO:0006511 ~ ubiquitin-dependent catabolic process)	(28)		(0.01)
Muscle	GO:0008104 ~ protein localization	84	2.7	0.003
Muscle	(GO:0046907 ~ intracellular transport)	(73)		(<0.001)
Muscle	GO:0005977 ~ glycogen metabolic process	8	1.6	0.009
Muscle	GO:0046324 ~ regulation of glucose import	10	1.5	<0.001
	Under-expressed in pigs from the low RFI line			
Muscle	GO:0006955 ~ immune response	62	10.1	<0.001
Muscle	GO:0002250 ~ adaptive immune response	18	5.4	<0.001
Muscle	GO:0001817 ~ regulation of cytokine production	22	4.6	<0.001
Muscle	GO:0010033 ~ response to organic substance	61	4	<0.001
Muscle	GO:0070482 ~ response to oxygen levels	17	4	<0.001
Muscle	GO:0001501 ~ skeletal system development	25	3.4	<0.001
Muscle	GO:0042981 ~ regulation of apoptosis	55	3	<0.001
Muscle	GO:0009743 ~ response to carbohydrate stimulus	10	2.1	<0.001
Muscle	GO:0006631 ~ fatty acid metabolic process	27	1.7	0.006
Muscle	(GO:0008610 ~ lipid biosynthetic process)	(19)		(0.05)
	Over-expressed in pigs from the low RFI line			
Liver	GO:0009100 ~ glycoprotein metabolic process	9	1.4	0.012
Liver	GO:0043691 ~ reverse cholesterol transport	3	1	0.024
Liver	GO:0007242 ~ intracellular signaling cascade	28	1	0.042
	Under-expressed in pigs from the low RFI line			
Liver	GO:0006952 ~ defense response	17	1.9	0.013
Liver	hsa00591:Linoleic acid metabolism	5	1.4	0.001
Liver	GO:0042981 ~ regulation of apoptosis	19	1.4	0.034
Liver	GO:0033559 ~ unsaturated fatty acid metabolic process	4	1.3	0.035
Liver	GO:0006468 ~ protein amino acid phosphorylation	16	1.3	0.049
Liver	GO:0050778 ~ positive regulation of immune response	7	1.1	0.016
Liver	GO:0006959 ~ humoral immune response	6	1	0.005
	Over-expressed in pigs from the low RFI line			
TASC	GO:0006796 ~ phosphate metabolic process	14	1.7	0.011
TASC	GO:0006006 ~ glucose metabolic process	5	1.6	0.018
TASC	GO:0006091 ~ generation of precursor metabolites and energy	11	1.5	0.004
TASC	(GO:0016042 ~ lipid catabolic process)	(6)		(0.006)
TASC	(GO:0006635 ~ fatty acid beta-oxidation)	(3)		(0.01)
TASC	GO:0034599 ~ cellular response to oxidative stress	4	1.2	0.033
	Under-expressed in pigs from the low RFI line			
TASC	GO:0007517 ~ muscle organ development	6	1.7	0.002
TASC	GO:0006796 ~ phosphate metabolic process	9	1.3	0.045
TASC	GO:0006952 ~ defense response	7	1.2	0.041
	Under-expressed in pigs from the low RFI line			
PRAT	GO:0001568 ~ blood vessel development	10	2.8	<0.001
PRAT	GO:0008202 ~ steroid metabolic process	7	1.4	0.012
PRAT	(GO:0008203 ~ cholesterol metabolic process)	(4)		(0.05)

Les principales conclusions de cette étude sont que la réponse immunitaire, la réponse au stress oxydatif et le métabolisme des protéines sont les principales voies biologiques communes aux quatre tissus qui distinguent les animaux des deux lignées. De nombreux gènes immunitaires sont sous-exprimés dans les quatre tissus des porcs de la lignée CMJR-, dont les gènes *CD40* (TNF receptor superfamily member 5), *CTSC* (cathepsin-C) et *NTN1* (Netrin 1). Différents gènes associés à l'utilisation de l'énergie semblent moduler de manière spécifique les tissus dans chaque lignée. Dans le muscle, la voie de régulation de l'utilisation du glycogène est surexprimée chez des porcs issus de la lignée CMJR- alors que les gènes impliqués dans le processus d'oxydation des acides gras sont sous-exprimés. A contrario dans les tissus adipeux cette voie est surexprimée chez ces mêmes animaux. Ce résultat indique que des stratégies opposées pour l'utilisation de l'énergie dans les muscles squelettiques et les tissus adipeux ont été mises en place dans chaque lignée. Enfin les gènes associés à la synthèse du cholestérol et aux flux dans le foie et les tissus adipeux sont également régulés de manières différentes chez les animaux de chacune des lignées.

Pour une majorité des processus biologiques mis en évidence dans cette étude, les résultats sont cohérents à la fois avec les caractéristiques phénotypiques connues des lignées divergentes du dispositif étudié mais aussi avec les résultats issus d'analyses d'animaux plus ou moins efficaces provenant d'autres dispositifs chez le porc ou chez d'autres espèces.

b. Analyses réalisées à l'aide de l'outil GSEA

Une nouvelle analyse d'étude d'enrichissement a été réalisée à l'aide de l'outil GSEA, prenant en entrée une liste de gènes ordonnés selon le différentiel d'expression mesuré sur chacun d'eux et cette liste a ensuite été confrontée à la base de données de termes GO humains filtrée sur les gènes uniquement retrouvés chez le porc. Pour cette approche nous avons donc récupéré l'ensemble des DEG identifiés par Gondret et al. (Gondret et al., 2017) pour les 4 tissus analysés et nous avons utilisé l'ensemble des paramètres par défaut de GSEA.

Au total, 693 termes GO ont été identifiés pour un seuil de FDR de 0,25 dont 91 avec une valeur de FDR inférieure à 0,01 et 149 compris entre 0,01 et 0,05. Parmi les 41 termes GO identifiés par Gondret et al à l'aide de l'outil DAVID, seul 12 termes sont retrouvés avec GSEA (Gondret et al., 2017). Néanmoins l'analyse des positions des termes dans l'arborescence de l'ensemble des GO indique que les GO identifiés dans les deux analyses sont fortement apparentés. La figure 44 illustre cette situation pour 3 des termes identifiés par Gondret et al dans la voie GO_0002376 (immune_system_process). Dans cette même voie, 35 termes ont été identifiés avec GSEA dont 11 avec une valeur de FDR <0,01 et 13 avec une valeur comprise entre 0,01 et 0,05.

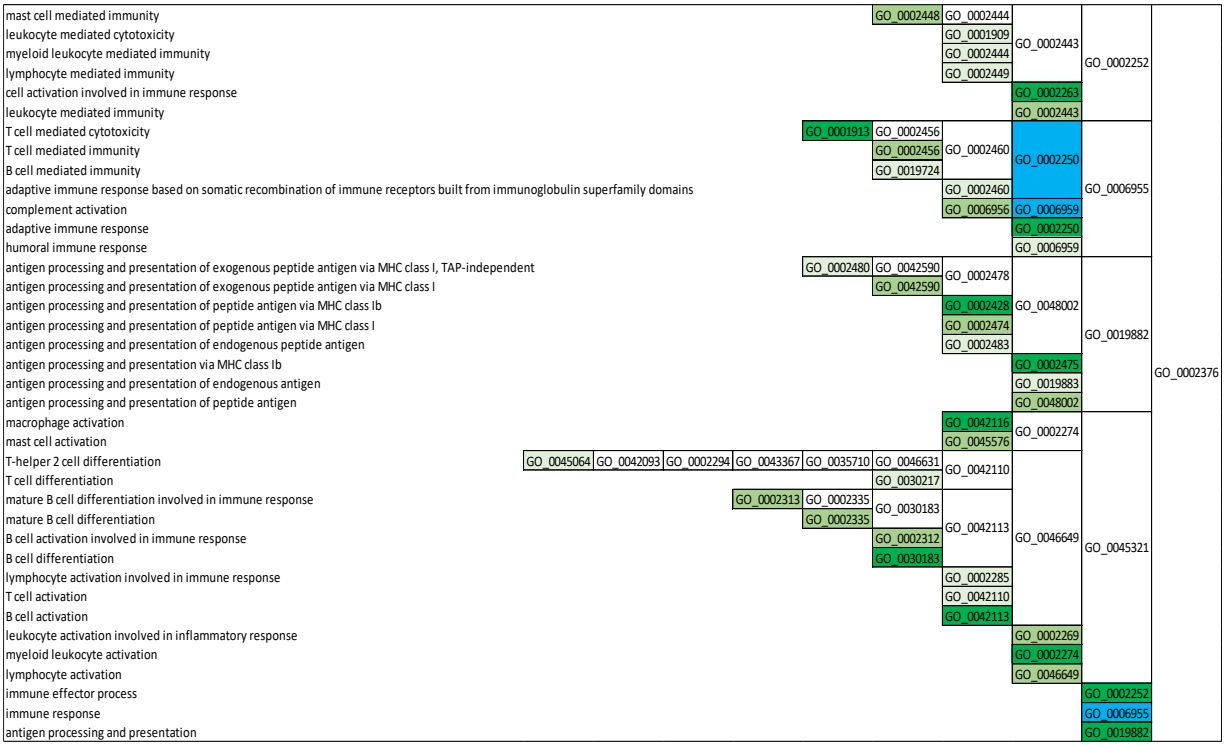

Figure 44 : Exemple d'apparentement entre les termes GO identifiés par Gondret et al. (en bleu) et les termes GO identifiés avec GSEA (en vert). Trois seuils de significativité sont indiqués, FDR<0,01 (vert foncé), 0,01 <FDR<0,05 (vert intermédiaire) et 0,05<FDR<0,25 (vert clair).

Dans la suite de ce chapitre, les résultats obtenus à chaque étape seront comparés à la liste complète de termes identifiés avec DAVID par Gondret et al. (Gondret et al., 2017) et les 681 termes GO complémentaires identifiés avec GSEA.

Au total, 12 501 gènes ont été analysés pour leur expression dans 4 tissus prélevés sur 48 animaux issus de la 8ème génération de sélection sur l'efficacité alimentaire. L'analyse transcriptomique a permis d'identifier 3 684 gènes différentiellement exprimés entre les lignées. A partir de cette liste de gènes une analyse d'enrichissement a permis de mettre en évidence 41 termes GO à l'aide de l'outil DAVID (Gondret et al., 2017) **et 681 termes significatifs pour un seuil de FDR de 0,25 via l'utilisation de GSEA.**

III. Combinaison des données génétiques et transcriptomiques

Cette seconde partie, est destinée à présenter les résultats obtenus en combinant la liste des DEG et les résultats d'analyses GWAS présentés dans le premier chapitre. Cette seconde approche permet ainsi de rechercher les gènes différentiellement exprimés (DEG) positionnés dans les régions QTL.

1. Les gènes différentiellement exprimés

a. Les gènes positionnés sur le génome du porc

Comme indiqué précédemment, à l'issu de l'analyse transcriptomique de 4 tissus appartenant à 48 animaux (24 animaux CMJR- et 24 animaux CMJR+), 3 684 DEG présentaient une valeur du différentiel d'expression supérieur à 1,1 pour les gènes surexprimés chez les animaux appartenant à la lignée CMJR- et inférieur à 0,91 pour les gènes surexprimés chez les animaux appartenant à la lignée CMJR+. Pour notre étude, il est nécessaire de connaître la position de chacun des gènes sur le génome du porc V11.1 ; cette information a donc été recherchée à l'aide du package *biomaRt* et de la fonction *getBM* sur R (V3.6.2). Parmi les 14 425 gènes présents sur la puce, 11 740 gènes au total sont positionnés sur les autosomes du génome du porc. En effet lors de la construction de la puce porcine, certaines sondes ont été choisies à partir de séquences de cDNA dont la position du gène correspondant n'est pas connue sur la séquence de référence porcine. Cette information manquante n'est pas contraignante pour des études transcriptomiques car l'information essentielle pour ces études est la fonction attribuée à chacun des gènes présents sur la puce. Au final pour la suite de notre analyse, seuls les gènes dont les positions sont connues sur la séquence du génome de porc, dans les bases de données Ensembl et du NCBI, ont été conservés soit 11 740 gènes sur le support dont 10 311 gènes sont exprimés dans l'étude de Gondret et al. (Gondret et al., 2017) et 3 122 différentiellement exprimés entre les lignées (Table 8). Afin de faciliter la combinaison des résultats des analyses transcriptomique et génétique, chacun de ces gènes a été assigné à une fenêtre de 1 Mb selon sa position sur le draft du génome de porc (V11.1) selon le même découpage du génome que celui réalisé pour l'interprétation des résultats de GWAS présentés dans le chapitre 1.

Table 8 : Présentation du nombre de gènes présents sur le support de la puce transcriptomique ainsi que le nombre de ces gènes positionnés sur le génome de porc V11.2. Au sein de ces deux groupes de gènes, deux catégories de gènes ont été prises en compte pour notre analyse : les gènes exprimés de la puce et parmi ces gènes ceux qui sont différentiellement exprimés entre les lignées (CMJR- et CMJR+).

Type de gènes	Nombre de gènes sur la puce	Nombre de gènes de la puce positionnés sur le génome de porc V11.2
Total	14 425	11 740
Exprimés	12 501	10 311
DEG	3 684	3 122

b. *Sélection des DEG les plus significatifs*

Parmi la liste de l'ensemble des DEG, nous avons fait le choix d'analyser plus spécifiquement les gènes présentant les plus forts différentiels entre les deux lignées : ce profil d'expression correspond à ce qui est attendu si une mutation affecte le promoteur d'un gène. Si ce gène est impliqué dans la variabilité du caractère CMJR, la comparaison du niveau d'expression du gène dans deux lots d'individus correspondant aux deux lignées divergentes présentera un différentiel fort. Nous avons choisi comme seuil un différentiel d'expression de 2. Au total 41 gènes ont été identifiés, correspondant à un différentiel d'expression ≥ 2 pour les gènes surexprimés en CMJR- et ≤ 0,5 pour les gènes surexprimés en CMJR+. La position des gènes de cette sous liste de DEG sur les autosomes du génome de porc sont présentés sur la figure 45.

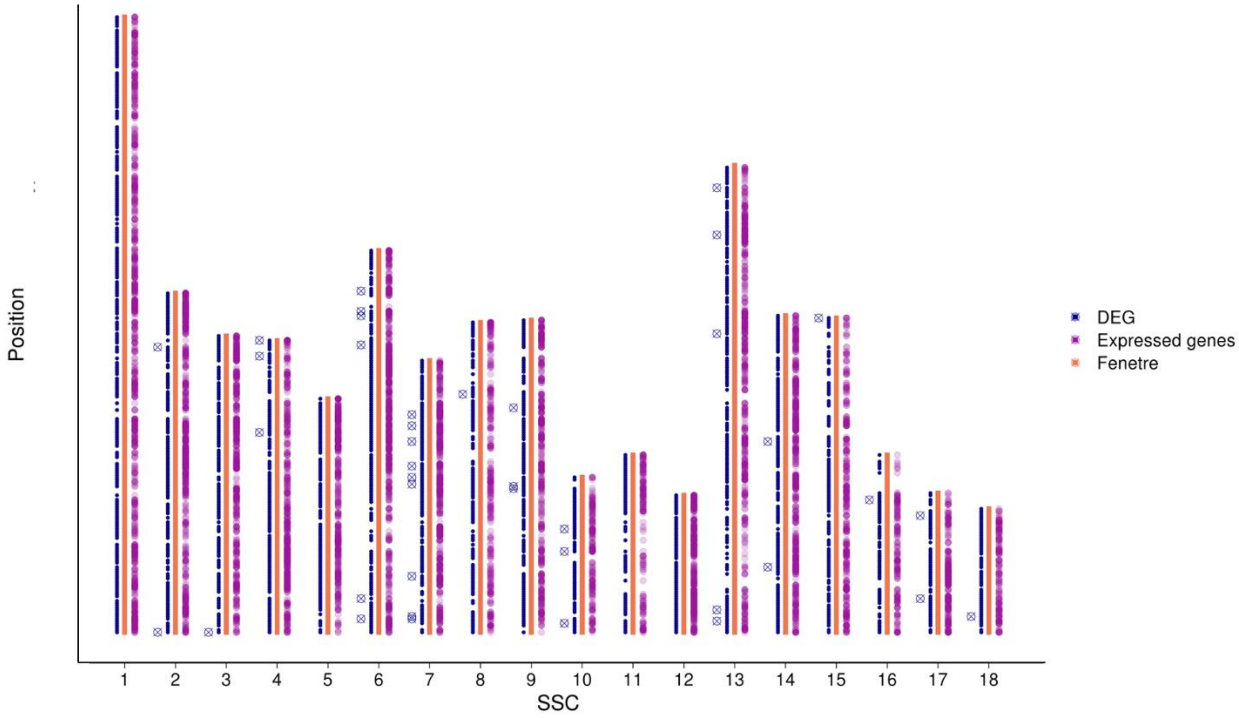

Figure 45 : Représentation sur le caryotype du génome de porc de la localisation de l'ensemble des gènes exprimés de la puce (pourpre) et des DEG (bleu) identifiés par Gondret et al. Les DEG avec un facteur de différentiel d'expression d'au moins 2 (dans l'une ou l'autre lignée) sont représentés sur le caryotype par des croix encerclées, à gauche des DEG.

2. Recherche de candidats positionnels dans les régions QTL parmi les DEG sélectionnés

Les analyses GWAS que nous avons réalisées nous ont permis de détecter 186 régions pour 24 caractères différents. Comme nous faisons l'hypothèse que le différentiel observé pour chaque gène résulte de la sélection sur le caractère CMJR nous avons donc choisi de ne garder que les régions QTL correspondant à la CMJR et aux caractères les plus corrélés avec la CMJR. Comme pour l'analyse des

évolutions de fréquences alléliques dans les régions QTL (Chapitre 1), nous avons choisi en plus de la CMJR, les 7 caractères présentant une corrélation génétique > |0,2| (selon les estimations réalisées par Gilbert et al. (Gondret et al., 2017) avec la CMJR : IC (0,39), temps d'imbibition du muscle (-0,29), pH 24h AD (0,28), IQV (0,26), CMJ (0,25), pH 24h GS (0,23) et pH 24h SM (0,22). Ces caractères correspondent à des caractères de mesure d'efficacité alimentaire et de qualité de viande, en revanche aucun caractère de poids de carcasse n'est corrélé avec la CMJR. Au lieu des 186 régions QTL, l'analyse sera faite avec 64 régions QTL réparties sur les 18 autosomes du génome de porc. En raison de la co-localisation de QTL détectés pour des caractères différents, ces 64 QTL sont localisés dans 48 régions uniques. Les régions QTL partagées le sont majoritairement entre des caractères de qualité de viande et certaines régions sont partagées entre un caractère de qualité de viande et un caractère d'efficacité alimentaire.

Nous avons alors estimé les distances les plus proches entre les 48 régions QTL et les fenêtres qui contiennent les 41 DEG présentant un différentiel d'expression supérieur à 2 et fait le choix de conserver les DEG localisés à moins de 3 Mb d'une fenêtre QTL. En effet le découpage en fenêtres de 1 Mb fixes est arbitraire et peut séparer une région QTL de DEG localisés à quelques kb. De plus compte tenu de la faible précision de localisation d'un QTL dans une analyse GWAS, il n'est pas exclu que le QTN recherché ne soit pas localisé dans la fenêtre contenant le SNP le plus significatif de l'analyse d'association. Au total, 9 fenêtres contenant des DEG forts ont été retenues et la liste de ces DEG et les caractéristiques des QTL auxquelles ils sont associés sont résumées dans la table 9.

Table 9 : Liste des DEG présentant un différentiel d'expression > 2 et localisés dans ou à proximité d'une région QTL.

Gene	Nom complet	SSC	Début_position	Fin_position	Num_Fenetre_DEG	Tissu	Distance au QTL le plus proche (Mb)	Intervel_QTL	Début_Fenetre_QTL	Fin_Fenetre_QTL	Caractère
ST3GAL1	CMP-N-acetylneuraminate-beta-galactosamide-alpha-2,3-sialyltransferase 1	4	7 814 694	7 909 528	568	LIVER	1	568_570	8 000 001	9 000 001	Ph24hAS_AFM
MT1A	Metallothionein-1A	6	18 660 642	18 673 766	815	PRAT	0	809_817	13 000 001	18 000 001	Ph24hAS_SM
CES3	Carboxylesterase 3	6	27 612 681	27 638 760	824	PRAT	1	822_823	26 000 001	27 000 001	IMB_time
CES1	Carboxylesterase 1	6	29 884 248	29 912 902	826	PRAT	3	822_823	26 000 001	27 000 001	IMB_time
CHGA	**Chromogranin-A**	**7**	**114 345 091**	**114 358 571**	**1082**	**PRAT**	**3**	**1084_1086**	**117 000 001**	**119 000 001**	**RFI_eq2**
SERPINA1	**Alpha-1-antitrypsin**	**7**	**115 604 285**	**115 614 633**	**1083**	**MUSCLE**	**2**	**1084_1086**	**117 000 001**	**119 000 001**	**RFI_eq2**
AKR1C1	Aldo-keto reductase family 1 member C1	10	65 576 464	65 678 248	1434	PRAT	0	1431_1437	63 000 001	69 000 001	IMB_time
AKR1C2	Aldo-keto reductase family 1 member C2	10	65 576 464	65 678 248	1434	PRAT	0	1431_1437	63 000 001	69 000 001	IMB_time
KCNIP2	**Kv channel-interacting protein 2**	**14**	**112 857 062**	**112 876 440**	**1902**	**MUSCLE**	**1**	**1900_1901**	**111 000 001**	**112 000 001**	**RFI_eq2**
ANK1	**Ankyrin-1**	**17**	**10 763 876**	**10 985 191**	**2163**	**LIVER**	**3**	**2159_2160**	**7 000 001**	**8 000 001**	**FCR_10w_S**
SLPI	**Antileukoproteinase**	**17**	**47 595 158**	**47 598 575**	**2200**	**MUSCLE**	**1**	**2200_2201**	**48 000 001**	**49 000 001**	**RFI_eq2**

Parmi les régions QTL comprenant un DEG fort, 3 régions ont été détectées pour le caractère CMJR et une pour le caractère IC. Une étude fonctionnelle de ces gènes par rapport aux données de la littérature a été réalisée et deux gènes avec un différentiel d'expression fort présentent des fonctions

intéressantes quant à la caractérisation des lignées : *CHGA* et *SLPI*. En effet, CHGA caractérisant la chromogranine, est une protéine membre de la famille des sécétogranines, c'est-à-dire des protéines sécrétoires neuroendocriniennes. Cette protéine est retrouvée dans les vésicules sécrétoires des neurones et des cellules endocrines, et son produit génique est un précurseur de différents peptides biologiquement actifs comme la vasostatine, la pancréastatine et la parastatine (Aunis and Metz-Boutigue, 2001). De plus, il a également été démontré que la CHGA produisait divers peptides bioactifs dont la pancréastatine (PST) qui a une activité dysglycémique. La PST régulerait donc le métabolisme du glucose, des lipides et des protéines dans le foie et les tissus adipeux. Chez les souris CHGA *knockout*, la PST induit la gluconéogenèse dans le foie (Valicherla et al., 2013). Ce type de processus biologique semble donc intéressant dans la manière de dégrader les composés issus de l'alimentation dans des organes que nous avons ciblés. D'autre part, *SLPI* est un gène qui code pour un inhibiteur sécrété protégeant les tissus épithéliaux des protéases à sérine. Ce gène a une affinité pour la trypsine, la chymotrypsine, l'élastase leucocytaire et la cathepsine G, et son effet inhibiteur contribue à la réponse immunitaire en protégeant les surfaces épithéliales contre l'attaque des enzymes protéolytiques endogènes. Cette protéine antimicrobienne a une activité antibactérienne de par sa modification des réponses inflammatoires et immunitaires après une infection bactérienne et après une infection par le parasite intracellulaire (Shin et al., 2019). L'implication de ce gène dans le phénotype des lignées peut-être en lien avec les différentes voies métaboliques de l'immunité reportées dans de précédentes études physiologiques sur les lignées.

Sur l'ensemble des DEG différentiellement exprimés, 9 DEG présentent un différentiel d'expression supérieur à deux entre les deux lignées et ont été identifiés dans ou à proximité des régions QTL. Parmi ces DEG, deux gènes localisés à proximité de deux des QTL identifiés pour le caractère CMJR présentent une fonction biologique intéressante qui pourrait être corrélée à des variations de l'efficacité alimentaire des animaux.

IV. Etudes d'enrichissement à partir des régions QTL

En faisant abstraction de la liste des DEG identifiés par Gondret et al. (Gondret et al., 2017), une approche alternative à la recherche de fonctions métaboliques impactées par la variabilité des gènes des régions QTL, est une étude d'enrichissement. Cette approche permet de cibler les gènes de régions d'intérêt du génome sans être dépendant de leur expression dans un tissu particulier ou un stade particulier. Différents outils sont disponibles pour réaliser des études d'enrichissement mais notre choix s'est porté sur l'outil GSEA (Subramanian et al., 2005)) qui présente l'avantage de pouvoir ordonner les gènes, caractérisés par des "informations génétiques", que l'on souhaite mettre en avant pour la détection de voies métaboliques spécifiques répertoriées dans le cadre du projet Gene Onthologie.

1. Les études d'enrichissement avec GSEA

a. Application de GSEA sur des résultats de GWAS

i. Les options de GSEA choisies

L'outil GSEA a initialement été développé pour l'identification de voies métaboliques caractéristiques des gènes différentiellement exprimés détectés lors de l'analyse de données transcriptomiques, comme nous l'avons utilisé dans la première sous-partie de ce chapitre. Cet outil permet d'ordonner les DEG selon la valeur du différentiel d'expression et cette ordonnancement a un rôle important dans l'identification des voies métaboliques significatives des DEG analysés. GSEA calcule différentes statistiques (ES, score d'enrichissement ; NES, score d'enrichissement normalisé ; p-value, ; FDR, *p*-value corrigée par le nombre de tests multiples) pour chaque voie métabolique identifiée grâce aux gènes à analyser. Dans le but d'identifier les voies métaboliques les plus spécifiques des données génétiques, l'option « classic » dans GSEA a été utilisée comme indiqué dans le manuel d'utilisation et discuté par Subramanian et al. (Subramanian et al., 2005), pour signifier lors des calculs statistiques sur les termes GO que les gènes analysés ne sont pas issus de données transcriptomiques. En complément du mode « classic », une valeur de pondération d'enrichissement *p* a été ajoutée dans le calcul de significativité des termes GO. Cette valeur d'enrichissement doit être choisie parmi 3 valeurs, soit une pondération de 1, 1,5 ou 2. D'après Subramanian et al. (Subramanian et al., 2005), une pondération de 1 correspond à l'étude de données transcriptomiques, or dans notre approche de détection de termes GO enrichis en gènes présents dans les régions QTL, nous avons privilégié la pondération maximale de 2 pour donner d'avantage de poids, dans l'analyse, aux gènes des régions QTL.

ii. Adaptation de la librairie de termes GO et de la liste à analyser

La librairie de termes GO : Les voies métaboliques analysées avec GSEA proviennent de la librairie de termes GO humain contenant les termes GO des composants cellulaires, des fonctions moléculaires et des processus biologiques. Au total, 9 996 termes GO sont présents dans la librairie avec en moyenne 85 gènes par terme GO. Dans notre étude de données issues de l'analyse du génome de porc, la librairie de termes GO humain a été filtrée sur tous les gènes positionnés sur le génome de porc dans le but d'obtenir les résultats statistiques les plus spécifiques et les plus précis par rapport à nos données. Avec ce filtre, aucun terme GO ne sera impacté dans le calcul statistique par l'absence, dans notre liste à analyser, des gènes non positionnés sur le génome de porc et donc ne pouvant pas être retrouvés dans les termes GO de la librairie. Cette étape de filtration de la librairie n'a pas engendré la perte d'un grand nombre de GO (7 termes GO). Au final, la librairie GO utilisée comporte 9 989 termes GO contenant en moyenne 72 gènes.

Les régions QTL sélectionnées : Trois analyses GWAS différentes ont été réalisées, des analyses globales et des analyses intra-lignées. Mais les résultats obtenus et décrits dans le chapitre 1 nous ont amené à conclure que la très grande majorité des régions détectées est impliquée dans l'architecture du caractère CMJR aussi bien dans la lignée à haute ou faible efficacité. Nous avons donc fait le choix d'utiliser l'ensemble des régions QTL identifiées pour les 3 analyses GWAS (Globale, CMJR- et CMJR+) dans une seule et même analyse GSEA. Les études d'enrichissement ont ensuite été réalisées pour les 24 caractères indépendamment, afin d'identifier les voies métaboliques spécifiques des résultats de GWAS de chaque caractère.

Critère d'ordonnancement de la liste de gènes: Les études d'enrichissement ont été réalisées en utilisant l'ensemble des gènes positionnés sur le génome de porc, comme détaillé dans le chapitre 1, et les résultats de GWAS ont été résumés dans des fenêtres de 1Mb caractérisés par le $-\log_{10}(p\text{-value})$ du SNP le plus fort de chaque fenêtre. Ce découpage du génome de porc a été conservé pour les études d'enrichissement. Les 17 283 gènes connus chez le porc ont donc été assignés à une fenêtre de 1Mb selon leur position sur le génome. A la différence des études d'enrichissement réalisées avec des données transcriptomiques, aucune valeur n'a directement été mesurée sur les gènes pris en compte, c'est pourquoi nous avons choisi de leur attribuer la valeur du $-\log_{10}(p\text{-value})$ de l'analyse GWAS, de la fenêtre dans laquelle le gène est localisé (Figure 46).

Gènes	Values	Fenêtres
GeneA	6.01	1323
GeneB	6.01	1323
GeneC	5.94	2
GeneD	5.88	1836
GeneE	5.3	959
GeneF	5.3	959

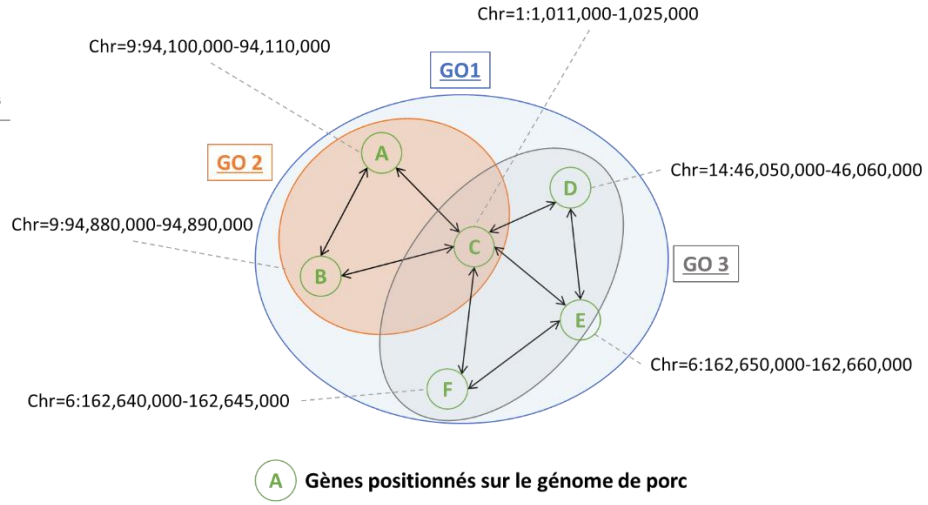

Figure 46 : Représentation schématique de le l'analyse d'enrichissement réalisé. Trois termes GO (1, 2 et 3) sont définis par une liste de gènes (A, B, C, D, E, F) positionnés sur différents chromosomes ou dans une même région génomique. Ces gènes sont alors ordonnés en fonction de la valeur du $-log_{10}$(p-value) attribué à la fenêtre où le gène est localisé.

Pour maximiser nos chances d'identifier tous les termes GO significatifs associés aux résultats de GWAS, deux méthodes d'ordonnancement des gènes ont été utilisés : (*i*) pour chaque fenêtre le $-\log_{10}$(*p*-value) **maximum** entre les 3 analyses GWAS, et (*ii*) la **somme** des $-\log_{10}$(*p*-value) obtenus dans les 3 analyses donnant dans ce cas plus de poids à des intervalles génomiques ayant été retrouvés significatifs dans plusieurs analyses GWAS comparés aux régions génomiques significatives que dans une seule analyse GWAS. Cette seconde méthode pour ordonner les gènes permet également de mieux prendre en compte les régions génomiques suggestives dans plusieurs analyses comparées à des régions significatives dans une seule analyse ou les régions sans association (ni significative, ni suggestive). A titre d'exemple, la comparaison des valeurs obtenues via ces deux approches pour le caractère CMJR est représentée à la figure 47. Le nuage de point obtenu reflète un ordonnancement différent des gènes dû aux valeurs utilisées. Deux études d'enrichissement avec deux listes contenant exactement les mêmes gènes mais ordonnés à l'aide de l'un ou l'autre des critères ont été réalisées pour les 24 caractères de consommation et d'efficacité alimentaire, de carcasse et de qualité de viande.

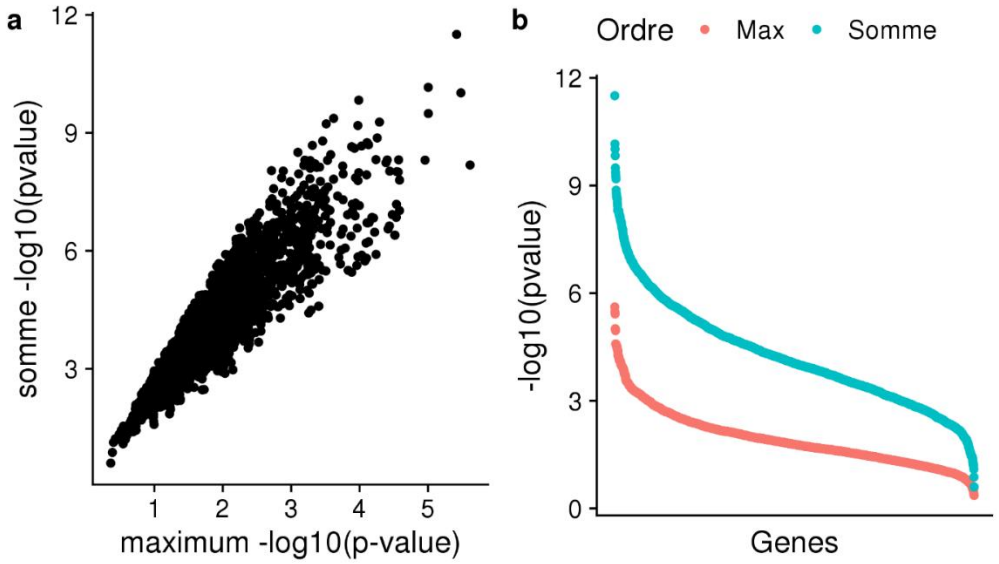

Figure 47 : Comparaison des deux valeurs d'ordonnancement : (a) distribution des maximum de -log10(p-value) (en rouge) et des somme des -log10(p-value) (en bleu). (a) Dot Plot pour l'ensemble des gènes de la valeur du maximum des -log10(p-value) versus la somme des -log10(p-value).

b. *Résultats obtenus à partir d'une liste de gènes*

Deux analyses GSEA ont été réalisées en parallèle en utilisant les deux valeurs d'ordonnancement pour l'ensemble des gènes positionnés sur le génome de porc, afin d'obtenir un score d'enrichissement pour la totalité des 9 989 termes GO présents dans la librairie filtrée pour les gènes connus chez le porc (Figure 48).

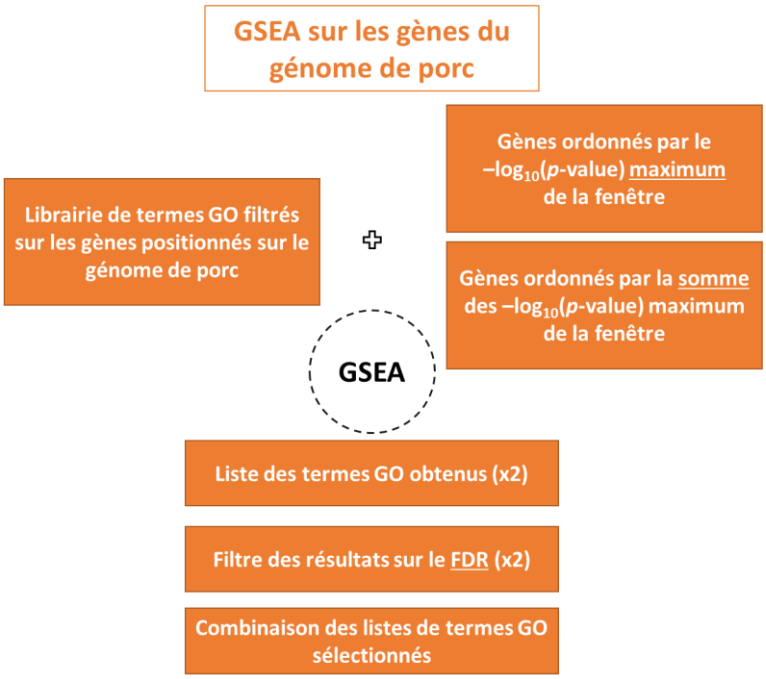

Figure 48 : Schéma de l'analyse GSEA effectuée sur la liste des gènes du génome de porc avec deux listes de gènes ordonnés soit par le maximum des –log10(p-value) de la fenêtre, soit par la somme des –log10(p-value) de la fenêtre. Deux listes de termes GO résultats sont obtenus.

Ainsi, deux listes de résultats de termes GO ont été obtenues et un seuil de 5% du FDR a été choisi pour filtrer les deux listes de termes GO. Au total, 28 termes GO ont été identifiés par les études d'enrichissement avec les gènes ordonnés par le $-\log_{10}(p\text{-value})$ maximal des fenêtres et 28 termes GO identifiés à partir de l'ordonnancement des gènes selon la somme des $-\log_{10}(p\text{-value})$ maximums des 3 analyses. Parmi ces deux listes, 15 termes GO sont communs. Au final, 41 termes GO uniques ont été sélectionnés grâce à cette approche (Table 10).

Table 10 : Table résultat de GSEA correspondant au 41 termes GO les plus significatifs avec le caractère pour lequel chaque GO présente la plus grande significativité. Seulement 3 valeurs statistiques issues de l'analyse GSEA sont représentées dans cette table : ES (Enrichment Score), NES (Normalize Enrichment Score) et FDR.q.val (False Discovery Rate).

NAME	Trait	ES	NES	FDR.q.val
GO_CALCIUM_DEPENDENT_CELL_CELL_ADHESION_VIA_PLASMA_MEMBRANE_CELL_ADHESION_MOLECULES	Mq_Index	0.7218167	2.1726658	0
GO_KERATINIZATION	Ham_W	0.73386735	2.2707796	0
GO_KERATIN_FILAMENT	Ham_W	0.84560007	2.274549	0
GO_CHEMOKINE_ACTIVITY	Shoulder_W	0.76173496	2.246964	0
GO_CORNIFICATION	Ham_W	0.7614132	2.323238	0
GO_EOSINOPHIL_MIGRATION	Belly_W	0.8196374	2.1857882	0
GO_CHEMOKINE_RECEPTOR_BINDING	Shoulder_W	0.7163193	2.1523354	0.0005040154
GO_CCR_CHEMOKINE_RECEPTOR_BINDING	Shoulder_W	0.73639154	2.1079092	0.0006723728
GO_EOSINOPHIL_CHEMOTAXIS	Shoulder_W	0.81856394	2.099537	0.00075669866
GO_CALCIUM_DEPENDENT_PHOSPHOLIPASE_A2_ACTIVITY	C_Lean_M	0.81946445	2.0606449	0.000997019
GO_THROMBIN_ACTIVATED_RECEPTOR_SIGNALING_PATHWAY	BF_W	0.86153346	2.077262	0.00197962
GO_NATURAL_KILLER_CELL_CHEMOTAXIS	Belly_W	0.897763	2.079794	0.002989506
GO_L_TYPE_VOLTAGE_GATED_CALCIUM_CHANNEL_COMPLEX	C_Lean_M	0.84745044	2.0278177	0.003972441
GO_PROTEIN_ARGININE_DEIMINASE_ACTIVITY	C_Lean_M	0.9889455	1.9856601	0.006359555
GO_HYDROLASE_ACTIVITY_ACTING_ON_CARBON_NITROGEN_BUT_NOT_PEPTIDE_BONDS_IN_LINEAR_AMIDINES	C_Lean_M	0.8316163	1.9935048	0.0064589363
GO_HISTONE_CITRULLINATION	C_Lean_M	0.9889455	1.9773391	0.007785647
GO_REGULATION_OF_RAC_PROTEIN_SIGNAL_TRANSDUCTION	b_GSM	0.81296563	2.070805	0.009010316
GO_CARBONATE_DEHYDRATASE_ACTIVITY	ADG_10w_S	0.8284105	2.013585	0.010022426
GO_CHEMOKINE_BINDING	Ham_W	0.7187617	1.9865333	0.0104786325
GO_INTERMEDIATE_FILAMENT_CYTOSKELETON	Ham_W	0.6179564	1.9303404	0.022322454
GO_PHOSPHATIDYLGLYCEROL_ACYL_CHAIN_REMODELING	C_Lean_M	0.74069446	1.930345	0.022580108
GO_REGULATION_OF_NATURAL_KILLER_CELL_CHEMOTAXIS	Belly_W	0.91088897	1.945097	0.023814108
GO_BUTYRATE_COA_LIGASE_ACTIVITY	Belly_W	0.964695	1.9494147	0.024421833
GO_POTASSIUM_ION_BINDING	Ham_W	0.8328098	1.9414219	0.025230825
GO_RECEPTOR_DIFFUSION_TRAPPING	C_Lean_M	0.8430636	1.9216393	0.026717832
GO_TRIGLYCERIDE_LIPASE_ACTIVITY	ABF_95kgBW	0.805946	1.9429804	0.026936145
GO_MODULATION_OF_GROWTH_OF_SYMBIONT_INVOLVED_IN_INTERACTION_WITH_HOST	Shoulder_W	0.7533486	1.995895	0.031027365
GO_CXCR_CHEMOKINE_RECEPTOR_BINDING	Shoulder_W	0.84445405	1.9576318	0.03133785
GO_PROTEASOME_ASSEMBLY	IMB_time	0.8138825	1.9936168	0.03243602
GO_RESPONSE_TO_CHEMOKINE	Shoulder_W	0.62012374	1.9635009	0.0328525
GO_INHIBITORY_EXTRACELLULAR_LIGAND_GATED_ION_CHANNEL_ACTIVITY	Ref_L_GMM	0.8141819	2.0264215	0.035005275
GO_PROTEIN_ACTIVATION_CASCADE	Ph24hAS_LDM	0.7399662	1.9966613	0.0381712
GO_ODORANT_BINDING	Ham_W	0.75166655	1.8847007	0.03925666
GO_RAGE_RECEPTOR_BINDING	Belly_W	0.85101074	1.9225658	0.039281055
GO_PHOSPHOLIPASE_A2_ACTIVITY_CONSUMING_1_2_DIPALMITOYLPHOSPHATIDYLCHOLINE_	C_Lean_M	0.7289242	1.8988492	0.039566625
GO_C_C_CHEMOKINE_BINDING	Ham_W	0.7000191	1.8793436	0.043983627
GO_STRESS_RESPONSE_TO_METAL_ION	Ph24hAS_LDM	0.94886565	2.005031	0.04523258
GO_NUCLEOBASE_METABOLIC_PROCESS	Ham_W	0.80098486	1.8937368	0.0455249
GO_THIAMINE_CONTAINING_COMPOUND_METABOLIC_PROCESS	Belly_W	0.92388713	1.8972181	0.04593254
GO_GLUCOCORTICOID_BIOSYNTHETIC_PROCESS	FCR_10w_S	0.72569907	1.9854348	0.04669842
GO_LYMPHOCYTE_CHEMOTAXIS	Belly_W	0.6210123	1.9013246	0.04826957

L'objectif de cette analyse est d'identifier si les gènes de certaines voies métaboliques sont particulièrement représentés dans les régions QTL. Afin d'affiner l'interprétation des résultats obtenus, les positions sur le génome des gènes des termes GO les plus significatifs ont été recherchées. Comme attendu le nombre de gènes par fenêtre est variable et certaines fenêtres ne contiennent qu'un seul gène alors que d'autres plusieurs dizaines. Lorsque les différents gènes d'une même fenêtre sont impliqués dans des voies métaboliques différentes la contribution de chaque fenêtre, intra terme GO, est la même. A contrario lorsque plusieurs gènes d'une même fenêtre sont impliqués dans un même terme GO, cette fenêtre contribuera plusieurs fois au score d'enrichissement du terme GO. Le score d'enrichissement sera d'autant plus impacté que ce cluster de gènes sera localisé dans une fenêtre QTL suggestive ou significative. Le terme GO "cornification" identifié à partir des gènes localisés dans les régions QTL mis en évidence pour le poids de jambon est une illustration de cette situation (Figure 49). Ce terme GO comprend 113 gènes dont 81 sont localisés dans 35 fenêtres sur le génome du porc. Ce terme est composé de plusieurs familles de gènes, la famille des Kératines (KRT) qui chez l'homme comprend 54 copies réparties en deux clusters sur les chromosomes 12 et 17, la famille des Kallikrein (KLK) regroupés chez l'homme en un cluster de 15 gènes sur le chromosome 19 et la famille des

Desmocollin (DSC) répartis entre deux clusters sur le chromosome 18. Chez le porc, ces clusters sont conservés et sont localisés dans les fenêtres, 912 (DSC), 852 (KLK), 709 et 1 540 (KRT). Cette situation est retrouvée pour la grande majorité des terme GO les plus significatifs de l'analyse.

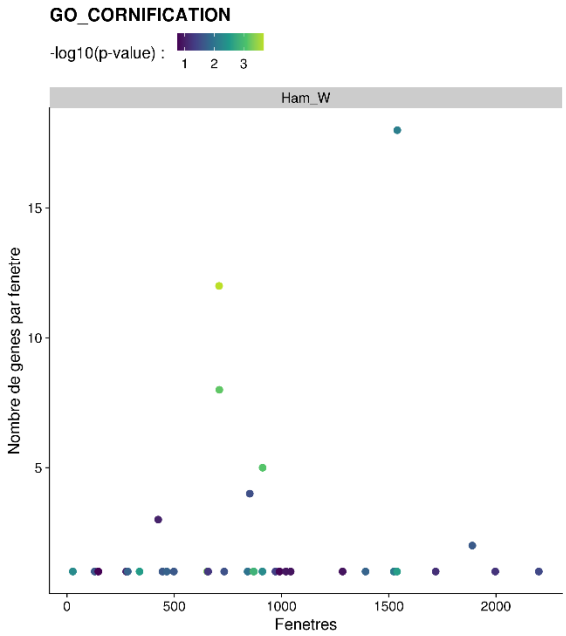

Figure 49 : Localisation le long du génome des fenêtres contenant les différents gènes du terme GO "cornification". Pour chaque fenêtre le nombre de gènes dans le GO est indiqué sur l'axe des Y. Le gradient de couleur indique la valeur du – log_{10}(p-value) maximum associé aux gènes de la fenêtre.

Les termes GO ciblés par notre approche ne doivent pas simplement résulter de la présence de nombreux gènes positionnés dans une même région génomique de 1 Mb et assignés à un même terme GO. La conclusion de cette analyse est que les clusters de gènes ont un poids trop important dans la significativité des termes GO lorsqu'ils se situent dans des régions QTL ou suggestives.

2. Une approche par fenêtre avec l'outil GSEA

Les résultats des études d'enrichissement sur les gènes nous ont amené à réfléchir à une stratégie alternative pour réaliser la détection de voies métaboliques significatives associées aux régions QTL. Les valeurs d'ordonnancement utilisées correspondent aux valeurs de p-values assignées aux 2 271 intervalles génomiques qui définissent le génome de porc V11.1. Nous avons fait le choix d'utiliser l'outil GSEA pour mener une étude d'enrichissement sur les intervalles génomiques et non plus sur les gènes.

a. *Les études d'enrichissement sur des fenêtres génomiques*

Comme résumé sur la figure 50, la première étape de cette analyse a été de transformer l'ensemble des termes GO de la librairie filtrée sur les gènes du génome de porc V11.1 afin de constituer des "termes de fenêtres".

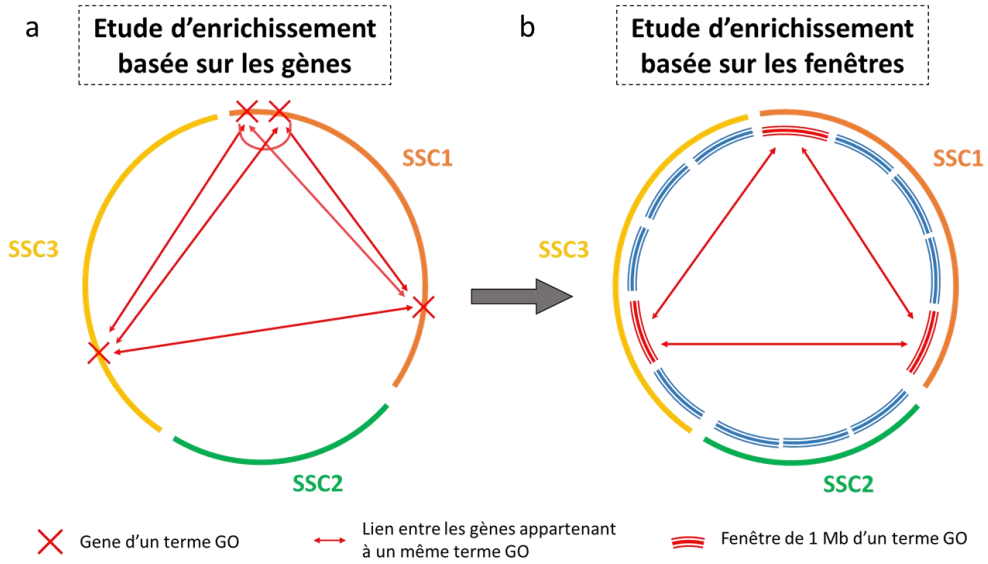

Figure 50 : Principe de transformation des termes GO caractérisant un ensemble de gènes possédant des caractéristiques fonctionnelles communes (a), en termes GO caractérisant des régions génomiques où sont positionnés des gènes avec des caractéristiques communes (b).

J'ai ainsi effectué le remplacement du nom de chaque gène par le numéro de la fenêtre (compris entre 1 et 2 271) dans laquelle il était localisé. A l'issu de ce remplacement, au sein de chaque terme GO, les fenêtres en doublon ont été supprimées afin que chaque terme GO ne comprenne que des numéros de fenêtre unique (Figure 51). Dans le cas des termes GO avec des clusters de gènes, une fenêtre unique remplace désormais une vingtaine de gènes. La conséquence majeure de cette approche est une diminution de la taille de certains termes GO, c'est-à-dire une diminution du nombre de fenêtres appartenant à ces voies métaboliques comparé au nombre de gènes les caractérisant. Au final le nombre de termes GO reste inchangé (9 989) mais le nombre d'éléments qui les composent est plus faible : 59 fenêtres par terme GO versus 85 gènes en moyenne dans la librairie initiale.

La seconde modification porte sur la liste soumise à l'analyse d'enrichissement. De manière équivalente, la liste est désormais composée des 2 271 fenêtres du génome ordonnées selon la valeur du $-\log_{10}$(p-value) des analyses GWAS (Figure 51).

Fenêtres	Values	Gènes
1323	6.01	GeneA/B
2	5.94	GeneC
1836	5.88	GeneD
959	5.3	GeneE/F

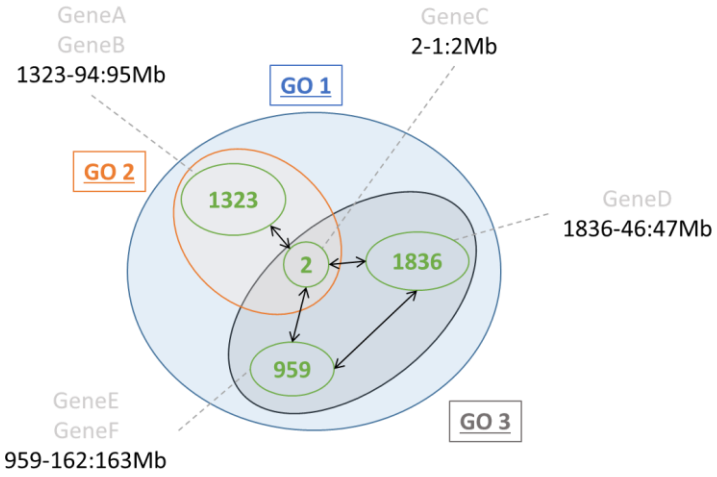

Figure 51 : Représentation schématique de le l'analyse d'enrichissement réalisé à partir des fenêtres. Le schéma est identique à celui présenté en figure 46 mais les gènes des termes Go et la liste soumise à l'analyse GSEA sont cette fois transformés en fenêtres ; les gènes localisés dans une même fenêtre (gènes A et B) n'apparaissent qu'une seule fois dans les termes GO et dans la liste ordonnée.

Hormis la transformation des gènes en fenêtres, les méthodes d'ordonnancement restent identiques aux études précédentes (partie IV.1.a). Deux listes de fenêtres ordonnées de manière différentes, soit en utilisant les valeurs de $-\log_{10}$(p-value) maximum parmi les 3 analyses GWAS soit en utilisant la sommes des $-\log_{10}$(p-value) maximum des 3 analyses, ont été analysées via GSEA.

b. *Identification des voies métaboliques caractéristiques du dispositif*

Les études d'enrichissement sur les intervalles génomiques du génome de porc ont donc été effectuées sur les 2 271 fenêtres et pour les 24 caractères indépendamment les uns des autres comme réalisé lors de l'analyse à partir des gènes (Figure 52). Cette nouvelle approche doit ainsi réduire le risque de biais créé par la présence de clusters de gènes, tout en maintenant possible la recherche de voies métaboliques surreprésentées dans les régions génomiques d'intérêts.

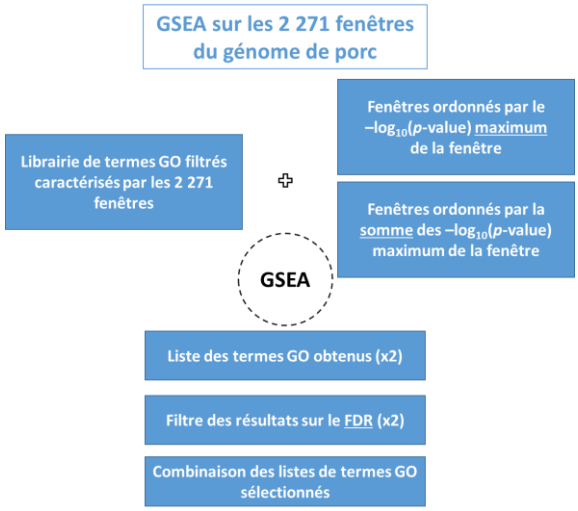

Figure 52 : Schéma de l'analyse GSEA effectuée sur la liste des fenêtres du génome de porc avec deux listes de gènes ordonnés soit par le maximum des –log10(p-value) de la fenêtre, soit par la somme des –log10(p-value) de la fenêtre. Deux listes de termes GO résultats sont obtenus.

Un seuil de FDR <25%, tel que recommandé dans le manuel d'utilisation de GSEA, a été appliqué et les listes de termes GO résultats issus des deux types d'ordonnancement ont été fusionnées pour chaque caractère. Au final pour l'ensemble des 24 caractères, 25 termes GO sont retrouvés significatifs à l'aide de l'ordonnancement des fenêtres par la somme des $-\log_{10}$(p-value) des 3 analyses GWAS. Les voies métaboliques sélectionnées sont présentées dans la table 11.

Table 11 : Liste des termes GO identifiés (FDR<0,25) avec les analyses GSEA effectuées sur les 2 271 fenêtres.

NAME	Trait	ES	NES	FDR.q.val
GO_POSITIVE_REGULATION_OF_VASCULAR_SMOOTH_MUSCLE_CELL_DIFFERENTIATION	Mq_Index	0.86860895	2.120092	0.01007227
GO_POSITIVE_REGULATION_OF_INSULIN_LIKE_GROWTH_FACTOR_RECEPTOR_SIGNALING_PATHWAY	Ref_L_GMM	0.8083227	2.0543303	0.01917705
GO_REGULATION_OF_AMYLOID_FIBRIL_FORMATION	b_GSM	0.8845399	2.0481095	0.02332255
GO_REGULATION_OF_MITOTIC_RECOMBINATION	b_GSM	0.921405	1.9909333	0.04760759
GO_POSITIVE_REGULATION_OF_BONE_RESORPTION	Ham_W	0.78094435	1.9836953	0.0614163
GO_GOLGI_TRANSPORT_COMPLEX	Ref_L_GSM	0.7921804	2.020816	0.07318619
GO_EOSINOPHIL_MIGRATION	Belly_W	0.7947267	2.0309587	0.0806045
GO_CARDIAC_CELL_FATE_COMMITMENT	Ph24hAS_GSM	0.7689568	2.012371	0.08413843
GO_WNT_SIGNALING_PATHWAY_INVOLVED_IN_HEART_DEVELOPMENT	Ph24hAS_GSM	0.8058534	1.9796199	0.09116414
GO_REGULATION_OF_RAC_PROTEIN_SIGNAL_TRANSDUCTION	b_GSM	0.7646185	1.9260516	0.12024654
GO_OXIDOREDUCTASE_ACTIVITY_ACTING_ON_PAIRED_DONORS_WITH_INCORPORATION_OR_REDUCTION_OF_MOLECULAR_OXYGEN_REDUCED_PTERIDINE_AS_ONE_DONOR_AND_INCORPORATION_OF_ONE_ATOM_OF_OXYGEN	b_GSM	0.87230563	1.9059786	0.12438218
GO_NEGATIVE_REGULATION_OF_TELOMERE_CAPPING	b_GSM	0.85725594	1.9345763	0.13467588
GO_NEURON_PROJECTION_MAINTENANCE	b_GSM	0.8097866	1.9085875	0.14237118
GO_POSITIVE_REGULATION_OF_ENDOTHELIAL_CELL_DIFFERENTIATION	Belly_W	0.7056487	1.9712148	0.14807704
GO_POSITIVE_REGULATION_OF_HELICASE_ACTIVITY	IMB_time	0.8791624	1.9322459	0.1621742
GO_INTERLEUKIN_23_PRODUCTION	b_GSM	0.81161296	1.8850315	0.16373855
GO_DEATH_RECEPTOR_BINDING	Ph24hAS_AFM	0.72422093	1.92648	0.16795631
GO_QUINONE_BIOSYNTHETIC_PROCESS	Mq_Index	0.7422679	1.9793962	0.18542928
GO_CELLULAR_RESPONSE_TO_LOW_DENSITY_LIPOPROTEIN_PARTICLE_STIMULUS	Ph24hAS_GSM	0.71140647	1.9034642	0.1865048
GO_LONG_CHAIN_FATTY_ACYL_COA_BIOSYNTHETIC_PROCESS	Ph24hAS_AFM	0.7077502	1.9355874	0.1878949
GO_NUCLEOTIDE_ACTIVATED_PROTEIN_KINASE_COMPLEX	a_GSM	0.76825786	1.9096819	0.20950435
GO_LINOLEIC_ACID_METABOLIC_PROCESS	Ph24hAS_AFM	0.80555457	1.9014329	0.2174543
GO_POSITIVE_REGULATION_OF_GRANULOCYTE_DIFFERENTIATION	Belly_W	0.8129685	1.8921573	0.2263529
GO_TRICUSPID_VALVE_DEVELOPMENT	Belly_W	0.8851551	1.9029602	0.24619997
GO_MICROTUBULE_SEVERING	IMB_time	0.80481404	1.8558534	0.24822715

Sur les 24 caractères étudiés, les termes identifiés sont associés aux poids de jambon et de panne et aux caractères de qualité de la viande (pH 24h adducteur, pH 24h fessier superficiel, IQV, temps d'imbibition du muscle, b* fessier superficiel, a* fessier superficiel, L* muscle fessier superficiel, L* muscle fessier moyen). Aucun terme n'a été obtenu pour les différentes mesures de l'efficacité alimentaire contrairement aux gènes identifiés dans la partie précédente. En cohérence avec ce résultat aucun des 11 gènes qui présentent des différentiels d'expression supérieurs à deux et localisés dans les régions QTL, ne sont présents parmi les termes identifiés ici.

Les résultats les plus intéressants et encourageants résultent de la comparaison des termes obtenus dans cette analyse aux termes identifiés via l'étude d'enrichissement faite à partir des données transcriptomiques seules (partie I). En effet, le seul point commun à ces deux analyses est l'utilisation des mêmes lignées, pour obtenir une liste de DEG (partie I) et une liste de fenêtre QTL (partie III). La comparaison des termes GO révèle qu'aucun n'est partagé mais dans plusieurs cas les termes fortement apparentés ont été identifiés dans les deux analyses comme les termes GO *CELLULAR_RESPONSE_TO_LOW_DENSITY_LIPOPROTEIN_PARTICLE_STIMULUS* ou *LINOLEIC_ACID_METABOLIC_PROCESS* (Figure 53).

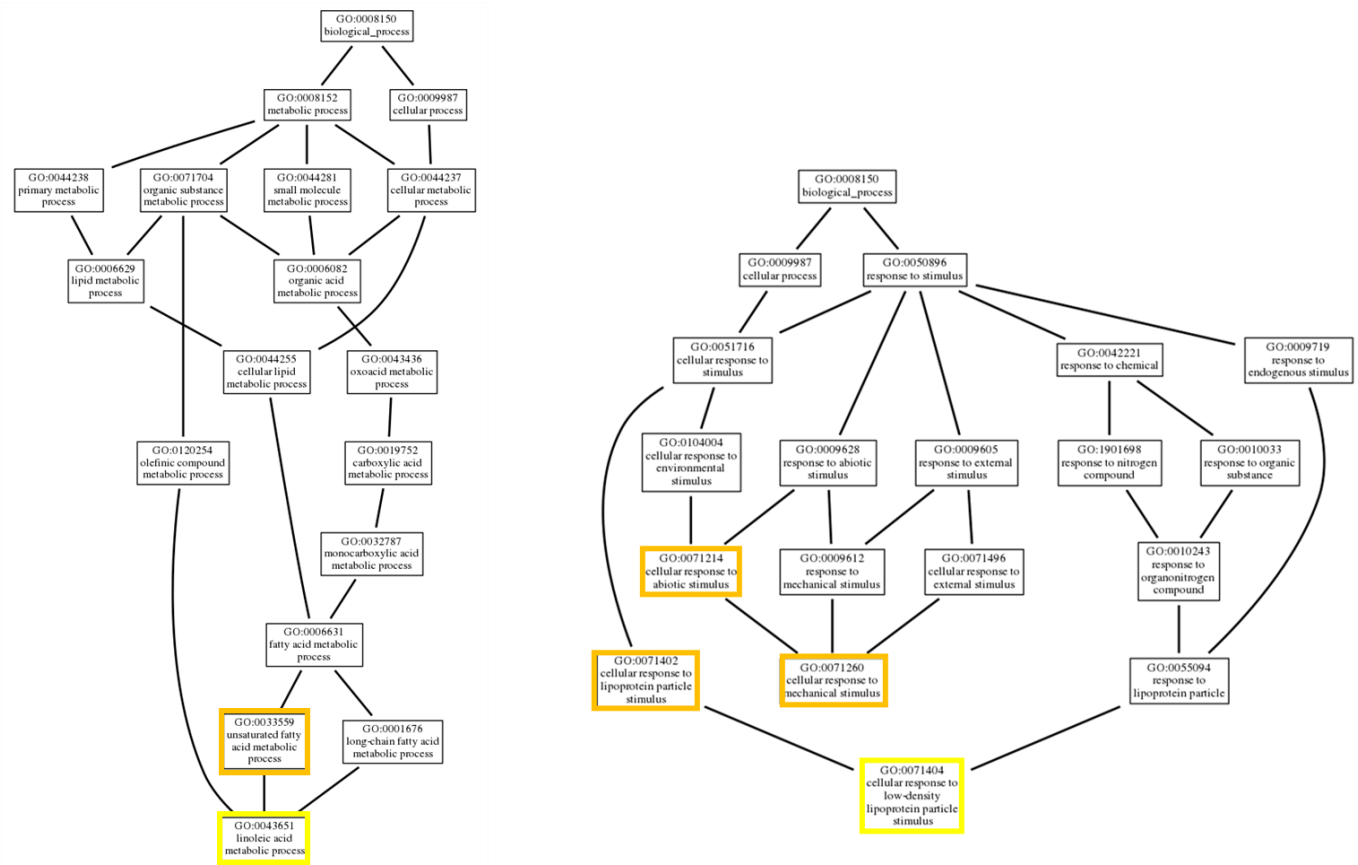

Figure 53 : Apparentement entre les termes GO identifiés à partir des fenêtres QTL (encadrés jaune) ou à partir de la liste de l'ensemble des DEG identifiés par Gondret & al. Dans les deux cas, les analyses ont été réalisées à l'aide de GSEA.

Enfin, pour chacun de ces termes la proportion de fenêtres qui présentent une valeur de $-\log_{10}$(p-value) d'au moins 3 a été évaluée. Cette proportion varie de 20% (*GO_POSITIVE_REGULATION_OF_BONE_RESORPTION*) à 66,7% (*GO_TRICUSPID_VALVE_DEVELOPMENT*). La proportion de fenêtres dont les valeurs de $-\log_{10}$(p-value) sont inférieures à 3 n'est pas surprenante car elle constitue la majorité des fenêtres de la liste soumise. Certains gènes de ces termes peuvent par exemple être localisés dans des fenêtres où l'informativité des SNPs ne permettait pas d'identifier des QTL.

Une analyse d'enrichissement à partir des fenêtres QTL a permis de mettre en évidence 25 termes GO. Certains de ces termes sont cohérents avec les résultats transcriptomiques obtenus dans la première partie de ce chapitre. L'objectif de la suite de cette analyse serait de pouvoir combiner simultanément dans une même analyse les données génétiques et transcriptomiques issues des lignées divergentes.

V. Analyse simultanée des 3 types de données

Jusqu'à présent les données disponibles ont été combinées selon trois stratégies : (*i*) une analyse d'enrichissement à partir d'une liste de DEG, (*ii*) une recherche dans les régions QTL des DEG dont l'expression diffère d'au moins un facteur 2 et (*iii*) une analyse d'enrichissement à partir d'une liste de gènes localisés dans les régions QTL. Dans cette dernière partie nous avons cherché à combiner l'ensemble de ces approches en réalisant une étude d'enrichissement à partir d'une liste de gènes différentiellement exprimés entre les deux lignées et localisés dans des régions sélectionnées sur la base d'informations génétiques.

De plus nous avons tenté dans cette dernière partie d'affiner nos critères de choix des DEG et des régions génétiques. En effet jusqu'à présent seules les régions QTL identifiées avec une analyse GWAS ont été prises en compte. Nous avons choisi de nous intéresser également aux régions du génome dont les fréquences alléliques ont fortement évolué entre les générations G0 et G7, régions candidates comme des "traces de sélection". Si ces régions contiennent des gènes contribuant à la variabilité de la CMJR, elles peuvent avoir été sélectionnées au cours des générations et présenter une structure génétique qui ne permette plus de les identifier via une étude d'association. Le second critère de choix que nous avons modifié concerne la sélection des DEG. Jusqu'à présent nous avons choisi d'utiliser l'ensemble des DEG (sous-partie 1) ou uniquement les DEG présentant un différentiel d'au moins 2 (sous-partie 2). De nombreuses études eQTL ont permis de mettre en évidence une large proportion de régulations en *cis*. Comme pour les analyses eQTL, nous avons fait l'hypothèse que certaines mutations génétiques (QTN) pourraient affecter un élément de régulation en *cis* qui affectera la régulation de plusieurs gènes successivement dans un même intervalle. La première étape de ce travail a donc été de sélectionner en complément des DEG localisés dans les régions QTL, ceux dont les profils d'expression présentaient des localisations en clusters.

En complément des fenêtres des QTL, et des fenêtres contenant les DEG présentant un différentiel d'expression d'au moins 2, nous avons recherché (*i*) les DEG présents en cluster dans des fenêtres de 1 Mb et (*ii*) sélectionné parmi ces fenêtres celles qui présentaient sur le plan génétique une structure apparentée à une trace de sélection. Une analyse d'enrichissement combinant ces différentes fenêtres a alors été réalisée.

1. Détection de clusters de DEG

Dans le cas où une mutation affecte la régulation en *cis* d'une région chromosomique, l'expression des gènes de l'ensemble de cette région sera affectée par cette mutation. Comme les régions génomiques où sont positionnés les 41 DEG avec un facteur 2 ont été étudiés de manière indépendante, ces gènes ne sont pas pris en compte dans la recherche de clusters de DEG. Au total 3 081 gènes avec un

différentiel d'expression significatif ont été assignés à une fenêtre de 1 Mb et le nombre de DEG dans chaque fenêtre d'1 Mb a ainsi pu être estimé (figure 54).

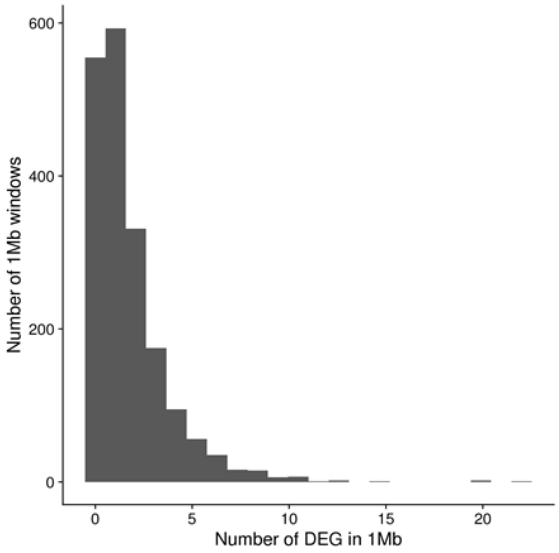

Figure 54 : Distribution du nombre de DEG par fenêtre de 1Mb pour les 2 271 fenêtres du génome de porc.

Au total, seules 555 fenêtres sur les 1 891 fenêtres du génome ne contiennent aucun DEG mais près de la moitié de ces fenêtres contenant des DEG ne contient qu'un gène (593 fenêtres) ou deux gènes (331 fenêtres) différentiellement exprimés. Notre objectif étant de détecter des clusters de gènes, nous avons fait le choix de nous focaliser principalement sur les fenêtres contenant au moins 3 DEG (967 fenêtres). Dans un second temps, deux filtres successifs ont été appliqués afin de prendre en compte le nombre de DEG et la moyenne de différentiel dans les fenêtres.

a. Proportion de DEG par fenêtre

En complément du nombre de DEG (au minimum 3) un critère de proportion a été estimé, car chaque fenêtre peut en effet contenir des gènes exprimés mais sans différentiel d'expression entre les lignées. J'ai donc calculé la proportion de DEG parmi le nombre de gènes exprimés dans la fenêtre, pour chaque fenêtre de 1Mb afin d'identifier les intervalles génomiques présentant un cluster de DEG important dans chaque intervalle contenant au moins 3 DEG. Les proportions de DEG dans les fenêtres de 1Mb sont comprises entre 14% et 100% (Figure 55). Notre objectif est d'identifier des fenêtres comprenant une proportion forte de DEG parmi l'ensemble des gènes exprimés dans l'intervalle afin de refléter une variation de la transcription en cluster. Nous avons choisi d'appliquer 4 seuils stricts différents en fonction du nombre de gènes compris dans chaque fenêtre : les fenêtres qui comprennent moins de 25% de DEG ont été écartées, les fenêtres présentant entre]25-50%] de DEG doivent comprendre au minimum 12 gènes, les fenêtres dont la proportion de DEG est comprise entre]50-75%] doivent

comprendre au minimum 6 gènes et les fenêtres comprenant de 3 à 6 gènes ne sont conservées que si elles comportent]75-100%] de DEG.

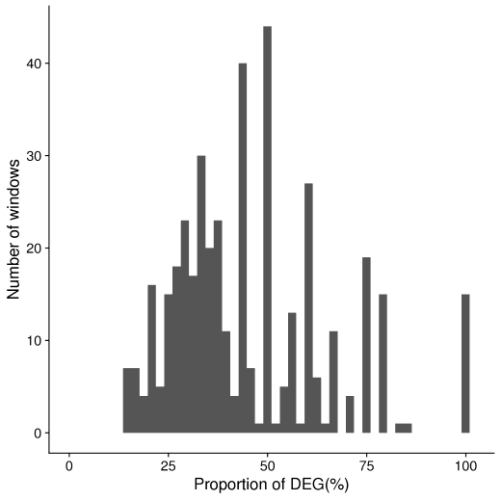

Figure 55 : Distribution de la proportion de DEG (en pourcentage, %) dans les fenêtres de 1 Mb contenant au moins 3 DEG.

Au final, 151 fenêtres ont été conservées et réparties dans les proportions suivantes : 99 fenêtres dans la proportion]25-50%], 22 fenêtres dans la proportion]50-75%] et 30 fenêtres dans la proportion]75-100%].

b. Moyenne d'expression par fenêtre

En complément de la prise en compte de la proportion de DEG et dans le but de sélectionner les fenêtres dont les clusters de DEG présentent les plus forts différentiels, la valeur moyenne du ratio d'expression des gènes de ces régions génomiques a été prise en compte. L'information sur la lignée dans laquelle chaque DEG est surexprimé n'étant pas pris en compte ici, l'inverse des valeurs de différentiels compris entre 0 < DEG < 0,9 correspondant aux gènes surexprimés dans la lignée CMJR+ ont été utilisés dans la moyenne afin que toutes les valeurs soient supérieures à 1,1. Les DEG dont la valeur du différentiel est supérieure à 2 (ou inférieure à 0,5) étant traités indépendamment, les moyennes par fenêtre sont bornées à 2. Pour le calcul de ces moyennes, les gènes exprimés (dont le ratio d'expression est égal à 1) et les DEG ont été pris en compte.

La distribution des valeurs moyennes des différentiels d'expression en fonction du nombre de gènes des fenêtres est représentée sur la figure 56. Il n'est pas surprenant de constater que les valeurs moyennes les plus fortes correspondent aux fenêtres comprenant le plus petit nombre de gènes (décroissance exponentielle). Comme pour la sélection des fenêtres présentant une forte proportion de DEG nous avons choisi d'appliquer un seuil par groupe de ratio d'expression. Trois seuils basés sur les moyennes globales du ratio d'expression dans chaque proportion ont été utilisés pour sélectionner

les fenêtres avec les différentiels les plus élevés de chaque proportion. Sur les 151 fenêtres sélectionnées sur la proportion de DEG, 71 fenêtres de 1 Mb ont été conservées sur la base de leur moyenne du différentiel d'expression des gènes les composant. A l'issue de cette sélection nous considérons que ces 71 fenêtres contiennent des clusters de gènes différentiellement exprimés.

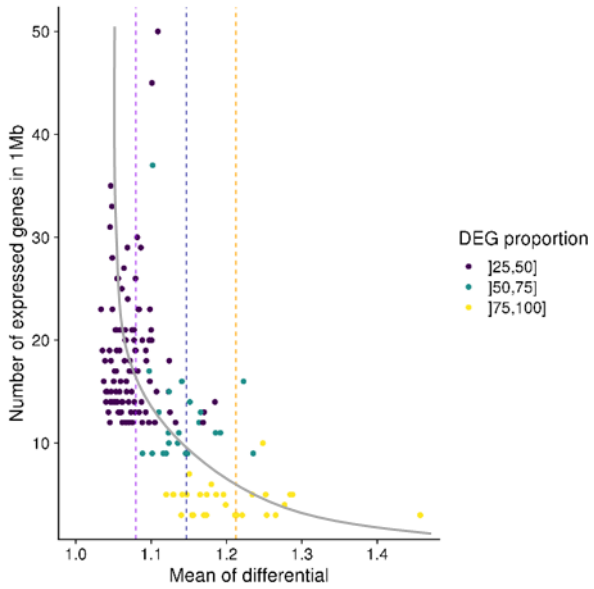

Figure 56 : Plot de la moyenne du différentiel par rapport au nombre de gènes dans la fenêtre pour les intervalles génomiques contenant au moins 12, 6 et 3 DEG dans les proportions de DEG]25-50] (violet),]50-75] (vert) et]75-100] (jaune) respectivement. En complément, pour chaque proportion, la moyenne globale des moyennes du différentiel de chaque fenêtre a été reportée via les lignes en pointillés de couleurs similaires à chaque proportion. Pour finir, la courbe représentant au mieux l'ensemble des points a été représentée en gris.

Sur l'ensemble des DEG différentiellement exprimés, 472 DEG localisés en clusters dans 71 fenêtres, ainsi que les 41 DEG présentant un ratio d'expression entre lignées d'au moins 2 (répartis dans 40 fenêtres) ont été conservés pour la suite de l'analyse.

2. Des régions génétiques "sous sélection"

Plusieurs régions QTL ont été détectées avec les GWAS, néanmoins les études d'associations ne sont jamais exhaustives et ne permettent pas de détecter l'ensemble des régions en ségrégation pour chaque caractère. Nous avons précédemment discuté le fait que les analyses GWAS réalisées sur chaque lignée peuvent manquer de puissance en fonction de la fréquence allélique des SNPs dans la région. De plus, nous sommes dans un dispositif familial pour lequel nous avons calculé un génotype moyen à partir des génotypes parentaux et l'ensemble des individus de chaque fratrie présente le même génotype. Nous sommes donc conscients d'une possible perte de puissance ne nous permettant pas de détecter toutes les régions variables comme significatives. En complément des régions QTL

détectées, nous nous sommes intéressés à l'identification de régions génomiques présentant un profil d'évolution de fréquences alléliques correspondant à un phénomène de sélection : il est en effet connu qu'au sein d'un isolat génétique, sous l'effet de la sélection d'un allèle "favorable" la fréquence de cet allèle augmente ainsi que les allèles en phase avec la mutation aux marqueurs SNP liés, par effet d'entrainement. La structure du déséquilibre de liaison (DL) dans la région sélectionnée tend ainsi également à augmenter. En complément des régions QTL nous avons ainsi choisi de rechercher et de sélectionner les régions du génome (dans les fenêtres de 1 Mb) présentant simultanément une forte évolution des fréquences alléliques entre les lignées et une évolution de la structure du DL (Figure 57).

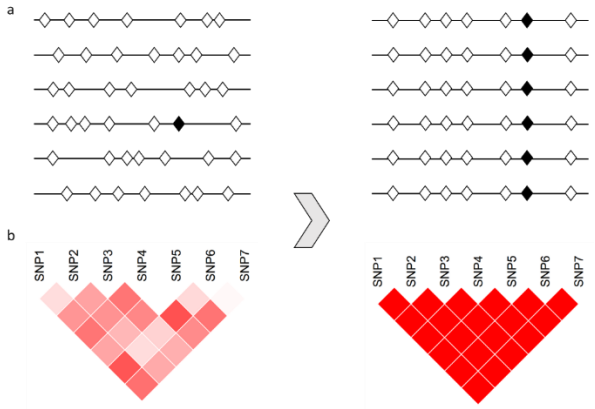

Figure 57 : Schéma de détection de régions en sélection par l'analyse des fréquences alléliques et du DL de manière commune. L'évolution des fréquences alléliques après plusieurs générations de sélection sont représentées dans la partie haute du schéma (a) et l'évolution du DL dans une région génomique en sélection est visualisée via des cartes triangulaires de chaleur.

Les régions génomiques identifiées à l'aide de cette approche, peuvent ainsi correspondre à des régions qui ont évolué dans une des deux lignées, ou à des régions dans lesquelles un haplotype différent a été sélectionné dans chaque lignée et qui n'avaient pas été détectées via l'analyse GWAS. L'objectif de notre étude étant de combiner les fenêtres des données transcriptomiques, nous avons fait le choix de limiter cette stratégie aux seules fenêtres sélectionnées pour leur contenu en DEG, à savoir les 40 fenêtres contenant un DEG de ratio d'expression supérieur à 2 et les 71 fenêtres contenant un cluster de DEG (103 fenêtres uniques).

a. Estimation de l'évolution des fréquences alléliques

L'analyse des fréquences alléliques a été menée via le calcul d'une moyenne par fenêtre des différences de fréquences alléliques aux SNPs de la fenêtre entre les animaux des lignées CMJR- et CMJR+. Les données transcriptomiques sont issues de tissus prélevés sur des animaux G8 descendants des reproducteurs G7 (P1). J'ai donc utilisé les génotypes des reproducteurs G7 pour étudier l'évolution

des fréquences alléliques. Les deux lignées étant issues d'une même population G0, l'évolution de fréquence sous l'effet de la sélection peut être résumée par la différence de fréquence entre les deux lignées en G7. Ces différences ont été calculées pour chaque marqueur et moyennées pour chaque fenêtre d'1 Mb. Les moyennes de différences de fréquences alléliques élevées correspondent à des fenêtres dont les SNPs présentent une évolution forte d'un allèle dans une lignée ou une sélection divergente des deux allèles dans chaque lignée (Figure 58a). En revanche, les fenêtres dans lesquelles les fréquences alléliques ont évolué dans la même direction pour les deux lignées, ou encore les régions dont les fréquences alléliques ont peu évolué dans les deux lignées au bout de 7 générations présentent des moyennes faibles (Figure 58b).

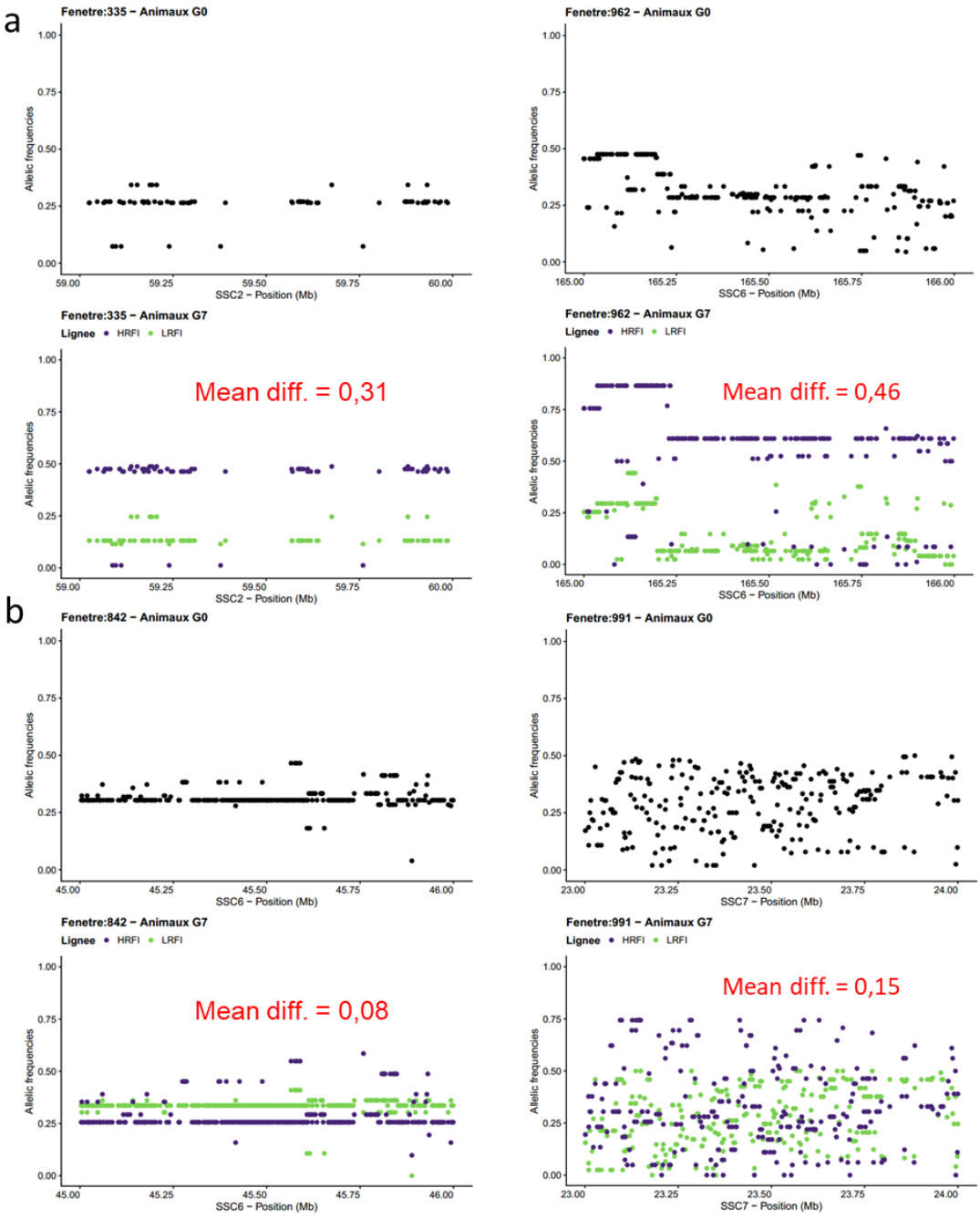

Figure 58 : Représentation de 4 profils différents d'évolution de fréquences alléliques entre les générations G0 et en G7 pour les deux lignées divergentes. Chaque paire de graphes représente une fenêtre d'1 Mb, et chaque point un SNP. Le graphe supérieur correspond aux fréquences d'un allèle de référence en G0, le graphe inférieur, la différence de fréquence entre les générations G0 et G7 pour chaque lignée (CMJR+ en violet, CMJR- en vert). La valeur moyenne de la différence des fréquences alléliques entre les lignées en G7 a été reportée sur chaque représentation en rouge. Les figures représentent (a) deux exemples de régions dont les fréquences ont fortement divergé entre les lignées et (b) deux exemples d'absence d'évolution entre lignées.

b. *Estimation de l'évolution de la structure du DL*

En parallèle de cette analyse des fréquences alléliques, une étude du DL a été menée afin d'estimer les fenêtres dont la structure a le plus évolué entre la G0 et la G7. Pour évaluer l'évolution du DL entre la

G0 et la G7 les matrices de corrélation (r^2) par paires de marqueurs hétérozygotes ont été calculées pour chaque génération (une valeur de r^2 =1 signifie un DL complet). Pour visualiser les résultats de l'analyse des blocs de DL, j'ai utilisé le package *gaston* V.1.5.7 sur R (V.3.6.2) pour réaliser des cartes triangulaires de chaleur du r^2 par paire de SNPs. Dans l'exemple présenté sur la figure 59, la structure du DL de la population G0 et de la lignée CMJR+ en G7 sont comparables. Les valeurs de r^2 sont légèrement supérieures mais c'est surtout dans la lignée CMJR- que deux sous-blocs se distinguent particulièrement à l'issue des 7 générations de sélection.

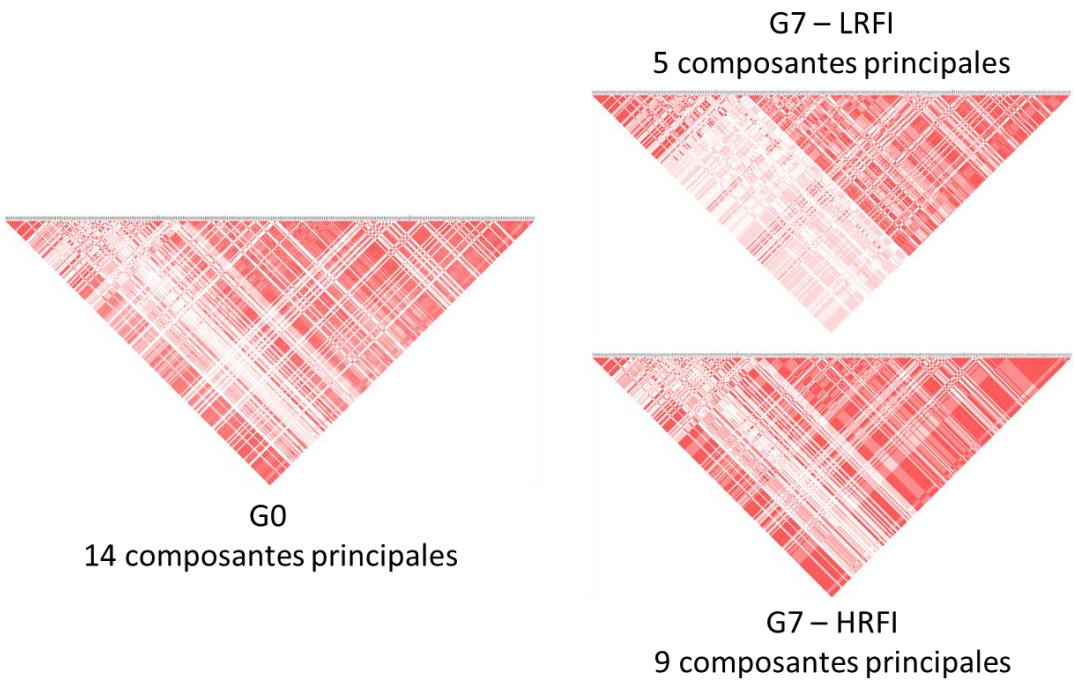

Figure 59 : Représentation des blocs de DL via des cartes triangulaires de chaleur du r^2, en G0 et G7 pour les 2 lignées (CMJR- et CMJR+) et pour un intervalle génomique donné. Le DL se visualise par un dégradé de rouge, directement lié aux valeurs de r^2 entre chaque marqueur de la région génomique (rouge : r^2 = 1; blanc : r^2 = 0).

Pour chacune des 103 fenêtres sélectionnées dans la partie IV.1 (présence d'un DEG avec un différentiel fort ou d'un cluster de DEG), la liste des marqueurs et leurs génotypes pour les individus G7 des deux lignées ont été sélectionnés pour calculer une matrice du DL par paires de marqueurs intra groupes d'animaux. Afin de résumer l'évolution du DL pour chaque fenêtre, j'ai choisi de calculer à partir d'une analyse en composantes principales (ACP) réalisée avec les valeurs de r^2 pour chaque fenêtre, le nombre de composantes principales représentant 99,5% de la variabilité génétique. Enfin pour estimer l'évolution de la structure du DL à partir de ce nombre de composantes nous avons calculé la proportion de composantes "perdues" entre les générations G0 et G7 : une simple différence du nombre de composantes entre les génotypes des fondateurs et ceux des individus de la G7 ne prend pas en compte le nombre de composantes identifiées en G0, or l'évolution faible du DL ne signifie pas

la même chose si la fenêtre de 1 Mb comporte un fort ou un faible DL dès la G0. C'est pourquoi nous avons choisi d'estimer la proportion maximale (valeur la plus forte entre CMJR+ et CMJR-) de composantes principales perdues entre la G0 et la G7. Cette proportion reflète ainsi l'évolution du DL sous l'effet de la sélection tout en tenant compte du nombre initial de composantes principales dans la région génomique étudiée.

La combinaison simultanée de l'évolution des fréquences alléliques et du déséquilibre de liaison permet de caractériser plusieurs profils de fenêtres (Figure 60) : (1) des fenêtres avec une faible différence des fréquences alléliques en G7 entre les lignées et une faible évolution du DL entre la G0 et G7 qui correspondent à l'absence de sélection, (2) des fenêtres avec une différence importante des fréquences alléliques entre les lignées en G7 mais toutefois une évolution du DL entre la G0 et la G7 très modérée qui peut être due à la présence d'un petit nombre d'haplotypes très structurés dès la génération G0 , (3) des fenêtres présentant d'importantes différences des fréquences alléliques entre les lignées en G7 et une forte évolution du DL entre la G0 et la G7 qui correspondent à la situation de régions sous sélection, et enfin (4) des fenêtres qui seraient caractérisées par une différence de fréquences alléliques très limitée en G7 entre les lignées CMJR- et CMJR+, mais la détection d'une évolution importante du DL entre la G0 et la G7 qui reflète une perte d'haplotypes sans sélection marquée d'un haplotype préférentiel.

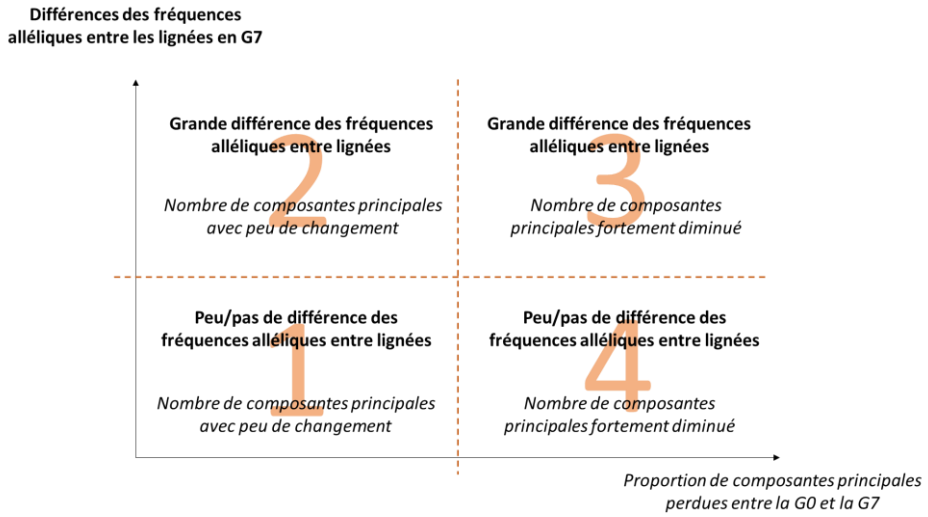

Figure 60 : Schéma des 4 profils d'évolution de la structure génétique des fenêtres sous l'effet de la sélection en fonction de la structure du DL et des fréquences alléliques aux SNP de la fenêtre.

Compte tenu de ces 4 profils d'évolution, seuls les profils 2 et 3 nous semblent intéressants dans la perspective d'identifier des régions génomiques en évolution au cours des générations et non identifiées à partir des études d'association. En effet, ces deux profils regroupent les fenêtres avec une

différence des fréquences alléliques et une structure de DL forte entre les lignées et ainsi ciblent les régions du génome dont les gènes seraient associés à la variabilité de la CMJR. En revanche, les fenêtres correspondant au profil 1 sont éliminées de l'analyse car caractéristiques de régions génomiques n'ayant pas évolué depuis 7 générations de sélection. Nous avons également fait le choix d'écarter les fenêtres du profil 4 car à l'issue de 7 générations de sélection malgré une augmentation du DL, peu de différences de fréquences alléliques existent entre les deux lignées.

Au vu de notre objectif, deux seuils ont été choisis pour identifier les différents profils d'évolution : une proportion de 50% de composantes principales perdues permet de dissocier les profils 1 et 2 des profils 3 et 4, et une moyenne de différence des fréquences alléliques de 0,3 pour différencier les profils 1 et 4 des profils 2 et 3 (Figure 61). Cette moyenne des fréquences alléliques a volontairement été choisie élevée pour ne garder que les fenêtres dont l'évolution est forte entre les lignées pour une majorité des SNP de la fenêtre. Les 103 fenêtres, comprenant un DEG à fort différentiel, ou un cluster de DEG se répartissent selon les 4 profils de la façon suivante : les profils 2 et 3 comprennent respectivement 3 et 13 fenêtres et les profils 1 et 4 comportent respectivement 18 et 69 fenêtres.

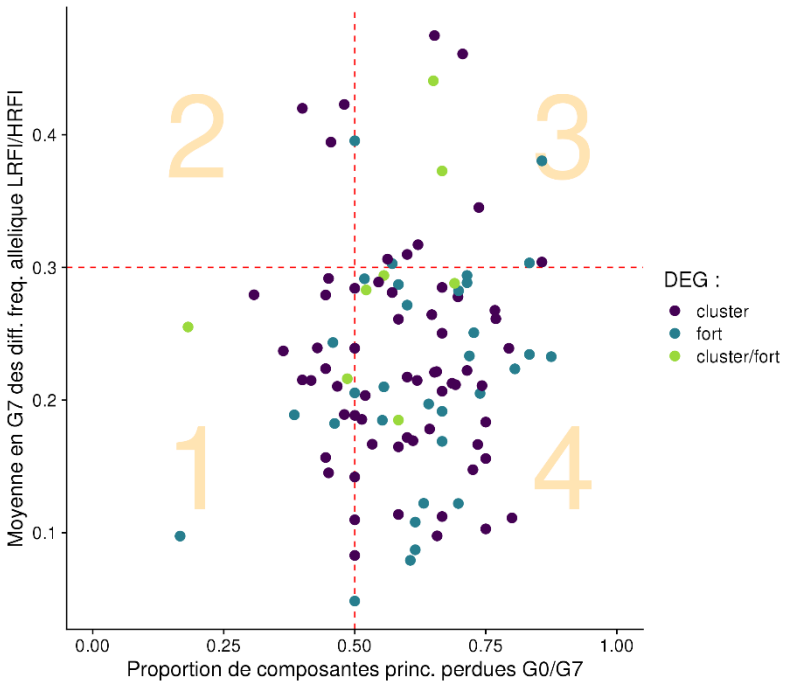

Figure 61 : Représentation de la proportion de composantes principales perdues entre la G0 et la G7, en fonction de la moyenne des différences alléliques en G7 entre les deux lignées pour les 103 fenêtres contenant les DEG sélectionnés (DEG avec différentiel > 2 et/ou des clusters de DEG). Les fenêtres violettes correspondent à celles possédant un cluster de DEG, les bleues sont celles contenant un DEG à différentiel fort et les vertes correspondent aux fenêtres contenant un DEG à différentiel fort et un cluster de DEG. Les pointillés rouges représentent les seuils choisis pour distingués les 4 profils d'évolution.

Au final, 16 fenêtres possèdent des profils d'évolutions alléliques correspondant à des profils de sélection entre les générations G0 et G7. Ces fenêtres sont aussi caractérisées par la présence d'un DEG avec un différentiel d'expression >2 et/ou un cluster de DEG.

3. Etude d'enrichissement

a. Liste des fenêtres sélectionnées pour l'analyse

La première comparaison a porté sur la recherche d'une co-localisation entre les régions QTL (48 régions identifiées pour les caractères CMJR et les caractères corrélés à la CMJR) et les fenêtres contenant les 41 DEG présentant un différentiel d'expression supérieur à 2 (identique à la sélection faite dans la sous-partie 2) ou les 71 fenêtres contenant un cluster de DEG. La distance entre les 103 fenêtres contenant les DEG (fort et/ou en cluster) et les 48 régions QTL sélectionnées, a été calculée. Au total, 7 fenêtres contenant des DEG forts et 10 fenêtres contenant un cluster de DEG ont été identifiées dans les régions QTL ou à quelques Mb. Deux fenêtres sont communes car elles contiennent parmi les DEG en cluster au moins un DEG dont le différentiel est supérieur à 2. La liste de ces fenêtres est rapportée dans la table 12.

Table 12 : Liste des fenêtres positionnées dans des régions QTL ou à quelques méga bases de celles-ci.

Fenêtre	SSC	# gènes	# DEG	# DEG fact. 2	Type DEG	CALCUL DISTANCE AU QTL LE PLUS PROCHE				
						Distance	Caractère(s)	QTL	Start (Mb)	End (Mb)
1434	10	9	2	2	DEG fact. 2	0	IMB_time	1431_1437	63	69
568	4	2	2	1	DEG fact. 2	1	Ph24hAS_AFM	568_570	8	9
824	6	22	7	1	DEG fact. 2	1	IMB_time	822_823	26	27
1902	14	12	4	1	DEG fact. 2	1	RFI_eq2	1900_1901	111	112
1083	7	7	4	1	DEG fact. 2	2	RFI_eq2	1084_1086	117	119
554	3	11	7	0	Cluster	0	DFI_10w_S	553_555	126	128
1201	8	3	3	0	Cluster	0	RFI_eq2	1199_1201	110	112
1296	9	13	5	0	Cluster	0	DFI_10w_S	1295_1296	67	68
1919	14	9	6	0	Cluster	0	RFI_eq2	1918_1920	130	131
2201	17	20	8	0	Cluster	0	RFI_eq2	2200_2201	48	49
266	1	13	4	0	Cluster	1	IMB_time	266_267	266	267
1225	8	12	6	0	Cluster	2	FCR_10w_S	1221_1223	132	134
2222	18	20	9	0	Cluster	2	IMB_time	2219_2220	4	5
815	6	10	6	1	Cluster & DEG fact. 2	0	Ph24hAS_SM	809_817	13	18
2200	17	14	7	1	Cluster & DEG fact. 2	1	RFI_eq2	2200_2201	48	49

Parmi les 88 fenêtres restantes, celles qui présentent un profil d'évolution génétique assimilable à des traces de sélection (profils 2 et 3) ont été sélectionnées. Les seuils appliqués ont ainsi permis la sélection de 13 fenêtres (Table 13) : 10 fenêtres contenant un cluster de DEG, 2 fenêtres possédant un DEG dont le ratio d'expression d'au moins 2 entre les lignées (« 602 » et « 1303 ») et 1 fenêtre partagée (présence d'un cluster de DEG et d'un DEG présentant un différentiel ≥ à 2 (« 1064 »)).

Table 13 : Liste des fenêtres avec évolution de structure génétique entre les lignées

Windows	SSC	# gènes	# DEG	# DEG facteur 2	Type DEG
335	2	26	8	0	Cluster
347	2	23	10	0	Cluster
962	6	12	7	0	Cluster
1026	7	23	8	0	Cluster
1417	10	8	5	0	Cluster
1537	12	21	9	0	Cluster
1539	12	29	13	0	Cluster
1573	12	5	4	0	Cluster
1831	14	13	6	0	Cluster
2186	17	7	4	0	Cluster
602	4	7	2	1	DEG facteur2
1303	9	5	4	1	DEG facteur2
1064	7	8	5	1	Cluster & DEG facteur2

b. *Analyses GSEA : intégration des données génétiques et transcriptomiques*

Comme pour les analyses présentées dans la sous-partie 3, les analyses GSEA ont été réalisées à partir des fenêtres et non des gènes. Les données soumises et la librairie de termes GO sont donc composées de numéro de fenêtres. Nous avons également choisi de conserver comme critère d'ordonnancement, le maximum ou la somme des valeurs de $-\log_{10}(p\text{-value})$ des résultats GWAS associés aux différentes fenêtres et de faire ainsi deux analyses indépendantes.

Pour cette analyse la difficulté rencontrée est que nous souhaitons que les fenêtres de clusters de gènes soient bien prises en compte dans l'estimation du score d'enrichissement et par conséquent soient présentes dans les premières positions de la liste soumise à GSEA. Or plusieurs de ces fenêtres ne co-localisent pas avec des fenêtres QTL et ne présentent donc pas des valeurs fortes pour leur ordonnancement. Nous avons donc fait le choix de leur attribuer une valeur par défaut de 4,5 (ordonnancement sur le maximum) ou de 10 (ordonnancement sur la somme), lorsque les valeurs respectives des $-\log_{10}(p\text{-value})$ étaient inférieures à ces seuils. A partir de ces nouvelles analyses 32

termes GO ont été identifiés avec un seuil de FDR < 0,25. La liste des termes est rapportée dans la table 14.

Table 14 : Liste des termes GO identifiés (FDR<0,25) avec les analyses GSEA effectuées sur les 2 271 fenêtres sélectionnées à partir des données génétiques et transcriptomiques.

NAME	Trait	ES	NES	FDR.q.val
GO_POSITIVE_REGULATION_OF_VASCULAR_SMOOTH_MUSCLE_CELL_DIFFERENTIATION	Mq_Index	0.8624198	2.069877	0.01923176
GO_NEGATIVE_REGULATION_OF_MULTICELLULAR_ORGANISM_GROWTH	Ph24hAS_AFM	0.8250009	2.005154	0.02950081
GO_COENZYME_A_BIOSYNTHETIC_PROCESS	Ph24hAS_AFM	0.8320416	2.012244	0.03620409
GO_DNA_N_GLYCOSYLASE_ACTIVITY	Ph24hAS_AFM	0.7748188	1.964574	0.05255259
GO_GTPASE_ACTIVATING_PROTEIN_BINDING	Ph24hAS_AFM	0.7853928	2.014808	0.06938995
GO_QUINONE_BIOSYNTHETIC_PROCESS	Mq_Index	0.736065	1.997613	0.08911058
GO_ANCHORED_COMPONENT_OF_SYNAPTIC_VESICLE_MEMBRANE	Ph24hAS_AFM	0.7573146	1.927514	0.09875885
GO_LONG_CHAIN_FATTY_ACYL_COA_BIOSYNTHETIC_PROCESS	Ph24hAS_AFM	0.7019827	1.900749	0.1103828
GO_BASE_EXCISION_REPAIR_AP_SITE_FORMATION	Ph24hAS_AFM	0.7937059	1.889591	0.11244454
GO_REGULATION_OF_METALLOPEPTIDASE_ACTIVITY	Ph24hAS_AFM	0.7548536	1.890293	0.12292696
GO_PEPTIDYL_ARGININE_METHYLATION	Ph24hAS_AFM	0.7362754	1.901669	0.12327755
GO_POSITIVE_REGULATION_OF_TRANSCRIPTION_BY_RNA_POLYMERASE_III	Ph24hAS_SM	0.8510229	1.989319	0.12478264
GO_POLYSACCHARIDE_BINDING	Ph24hAS_AFM	0.6977918	1.906771	0.12823492
GO_DEATH_RECEPTOR_BINDING	Ph24hAS_AFM	0.7189041	1.858074	0.16223365
GO_SNORNA_3_END_PROCESSING	Ph24hAS_AFM	0.6979077	1.858115	0.17566904
GO_WNT_SIGNALING_PATHWAY_INVOLVED_IN_HEART_DEVELOPMENT	Ph24hAS_GSM	0.7987769	1.980857	0.17611165
GO_ALPHA_LINOLENIC_ACID_METABOLIC_PROCESS	Ph24hAS_AFM	0.7605279	1.860992	0.18130784
GO_CELLULAR_RESPONSE_TO_COPPER_ION	IMB_time	0.7678323	1.871924	0.18211374
GO_GLIAL_CELL_PROJECTION	IMB_time	0.7282497	1.857035	0.18642893
GO_LINOLEIC_ACID_METABOLIC_PROCESS	Ph24hAS_AFM	0.7993681	1.841806	0.18914291
GO_ASTROCYTE_END_FOOT	Ph24hAS_AFM	0.9052078	1.843266	0.19798401
GO_STEROL_ESTERASE_ACTIVITY	Ph24hAS_AFM	0.9120331	1.835894	0.19837725
GO_CARDIAC_CELL_FATE_COMMITMENT	Ph24hAS_GSM	0.761874	1.945071	0.1998044
GO_REGULATION_OF_TRANSCRIPTION_BY_RNA_POLYMERASE_III	Ph24hAS_SM	0.7037407	1.9375	0.20382273
GO_MICROTUBULE_SEVERING	IMB_time	0.7995067	1.812442	0.20416671
GO_RESPONSE_TO_INTERFERON_ALPHA	IMB_time	0.744961	1.814471	0.21787974
GO_BLEB	IMB_time	0.8138982	1.819086	0.22194082
GO_ASTROCYTE_PROJECTION	IMB_time	0.8137381	1.939384	0.22358608
GO_POSITIVE_REGULATION_OF_HELICASE_ACTIVITY	IMB_time	0.8729759	1.905273	0.23328896
GO_REGULATION_OF_HELICASE_ACTIVITY	IMB_time	0.7650727	1.830117	0.235939
GO_REGULATION_OF_MICROTUBULE_DEPOLYMERIZATION	IMB_time	0.701044	1.836923	0.2422457
GO_MULTI_CILIATED_EPITHELIAL_CELL_DIFFERENTIATION	Ph24hAS_SM	0.8585516	1.866211	0.24320546

Entre les listes de termes GO obtenus via cette analyse ou uniquement à partir de la liste des régions QTL (sous-partie 3), 8 termes sont partagés et sont dans les deux cas uniquement identifiés grâce à des gènes localisés dans les régions QTL. Pour les 24 nouveaux termes, les caractères quantitatifs associés sont de nouveaux les phénotypes liés à la qualité de la viande. Aucun terme n'a été identifié pour les mesures associées à l'efficacité alimentaire.

Par rapport à l'analyse d'enrichissement réalisée dans la partie III de ce chapitre, les résultats attendus via cette nouvelle analyse sont de deux natures : (*i*) l'identification de nouveaux termes composés de gènes localisés dans les régions sous sélection et les régions QTL, et non identifiés en ne prenant en compte que les régions QTL (partie III), et (*ii*) l'identification de termes fortement associés

à ceux identifiés précédemment. La première situation correspond à ce que nous avons trouvé à partir de 4 termes GO associés au caractère temps d'imbibition du muscle (3 termes) et pH 24h adducteur (1 terme). Ces 4 termes GO présentent une forte proximité (Figure 62).

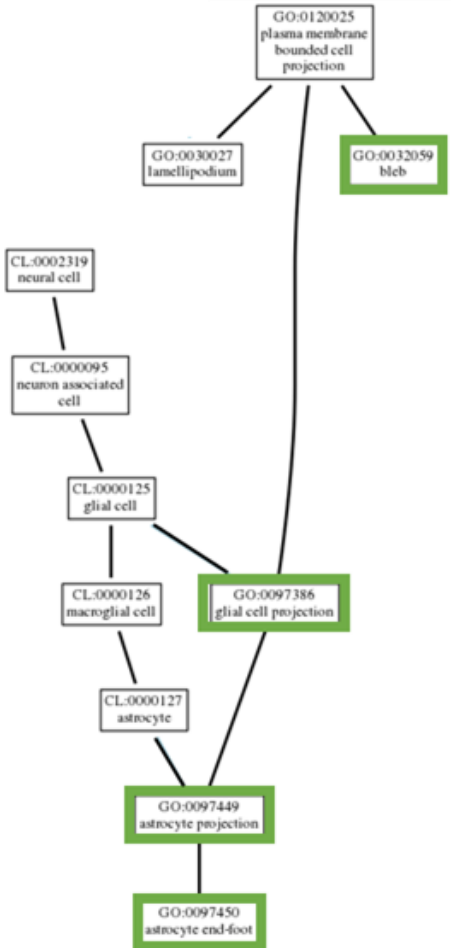

Figure 62 : Apparentement entre 4 termes GO identifiés pour les caractères du temps d'imbibition du muscle et du pH 24h adducteur

Au total ces 4 termes GO comprennent 31 gènes et parmi nos données 6 sont localisés dans des régions QTL suggestives, 4 dans des régions QTL significatives, 3 sont localisées dans des fenêtres dont la fréquence a fortement évolué au cours des générations. La seconde situation correspond au cas illustré dans la figure 63.

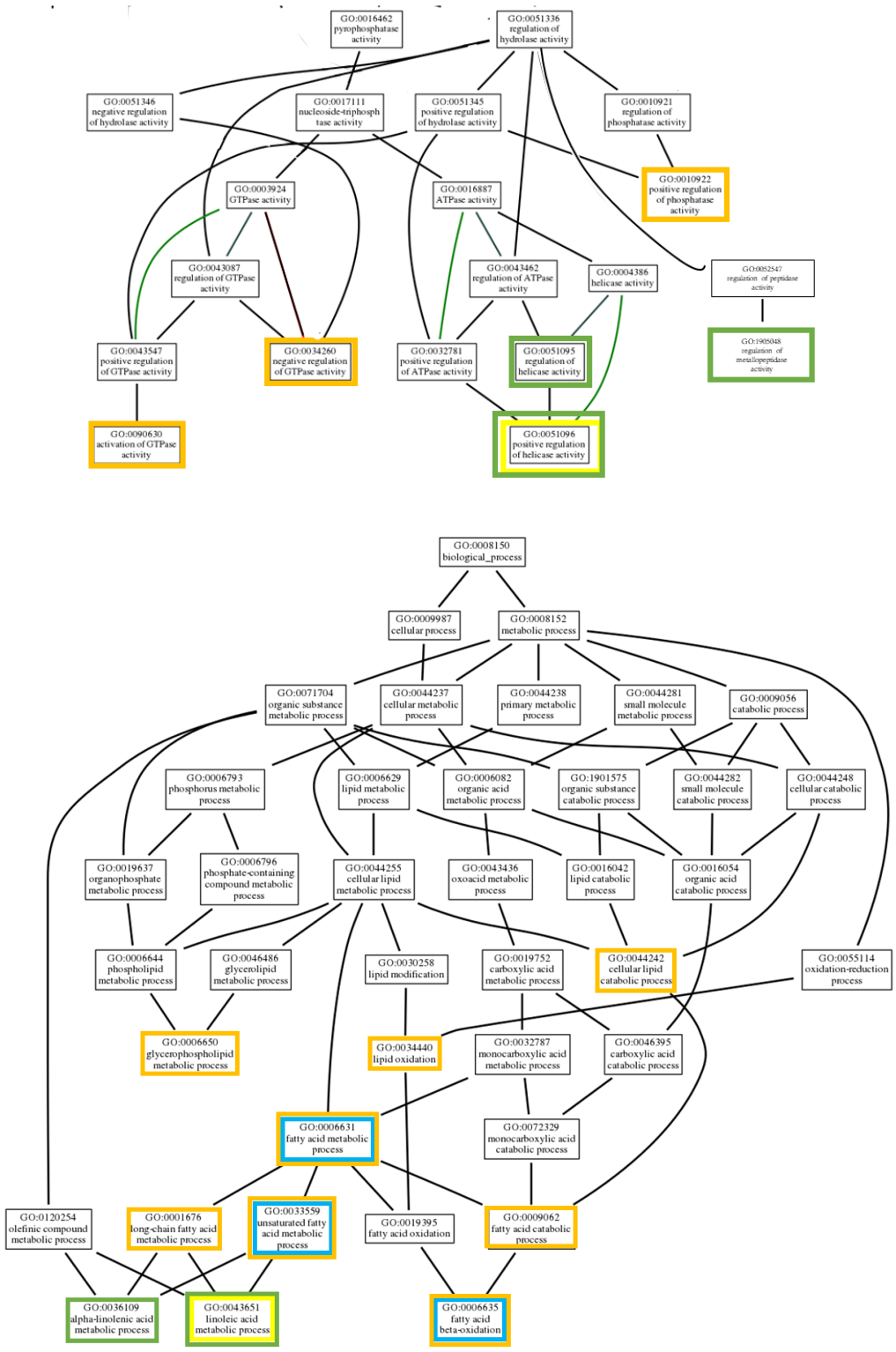

Figure 63 : Apparentement entre les termes GO identifiés dans les analyses successives : en bleu les termes GO identifiés par Gondret et al. (Gondret et al., 2017) à l'aide de l'outil DAVID, en orange les termes GO identifiés à partir de la même liste de DEG et l'outil GSEA (partie I), en jaune les termes GO identifiés à partir des régions QTL (partie II) et en vert les termes GO identifiés en combinant les informations génomiques et transcriptomiques (partie IV).

VI. Discussion

Ce chapitre présente diverses approches d'intégration de données basées à la fois sur des données génétiques et phénotypiques provenant du dispositif expérimental analysées via les analyses GWAS du chapitre 1, un jeu de données également obtenu sur le dispositif mais ne faisant pas parti initialement de mon projet et des bases de données publiques. Cette approche a été menée pas à pas afin d'évaluer l'apport respectif des données de nature et d'origine différentes. Les résultats les plus encourageants ont été d'obtenir des voies métaboliques proches via des études d'enrichissement à partir de la liste de DEG issue d'une analyse transcriptomique et via des études d'enrichissement à partir d'une liste de gènes localisés dans les régions QTL. La combinaison de l'ensemble des données n'a malheureusement pas apporté les résultats escomptés en raison de difficultés méthodologiques de combinaison de données issues d'analyses de natures différentes. Cette discussion est destinée à faire le point sur les résultats encourageants et à proposer des pistes de réflexion à explorer dans la suite de ce travail.

1. Vers une meilleure compréhension fonctionnelle des lignées divergentes ?

Une approche innovante a été réalisée avec les logiciel GSEA pour l'analyse des données génétiques via des études d'enrichissement. En effet, GSEA avait principalement été développé pour les études d'enrichissement (Subramanian et al., 2005) à partir de données fonctionnelles mais l'identification de régions QTL à partir des données génétiques nous a incité à réaliser des études d'enrichissement à partir des gènes de ces régions, en faisant l'hypothèse que les différents QTL impliqués dans le déterminisme de la variabilité d'un caractère pouvaient appartenir à une même voie métabolique. Parmi l'ensemble des voies biologiques identifiées, la voie impliquée dans la différenciation cellulaire du muscle a été mise en évidence dans mon approche mais avait également été reportée par Vincent et al. (Vincent et al., 2015). Dans cette même étude, les données transcriptomiques et protéomiques ont permis de mettre en évidence la voie métabolique des lipides, or dans notre étude, nous avons détecté une voie métabolique en lien avec des stimulus des lipoprotéines, molécules chargées du transport des lipides dans le sang. La voie métabolique des lipides a également été reportée lors de l'analyse transcriptomique des lignées divergentes de l'ISU (Liu et al., 2016) et dans celle des lignées INRAE (Gondret et al., 2017; Jégou et al., 2016). Enfin le terme GO de la régulation positive de l'insuline semble être une voie très intéressante à considérer. Cette voie métabolique n'a pas été identifiée par les études transcriptomiques issues du dispositif ISU mais Liu et al. (Liu et al., 2016) ont mis en évidence via la base de données IPAD (Integrated Pathway Analysis Database) (Zhang and Drabier, 2012) un enrichissement en gènes dans la voie de régulation de l'insuline, et notamment associées à des maladies de résistance à l'insuline.

Néanmoins certaines voies métaboliques identifiées ne sont pas partagées à l'issu des différentes analyses réalisées et certains mécanismes biologiques mis en évidence semblent peu explicatifs des phénotypes étudiés. Ces différences peuvent aussi être liées aux bases de données interrogées par chacun des outils permettant la réalisation des études d'enrichissement. En effet la multitude de logiciels et d'outils permettant de réaliser ce type d'approche possède d'une part des méthodes statistiques différentes mais également des bases de données différentes. L'outil GSEA est basé sur la base de données MSigDB (Molecular Signatures Database) (Liberzon et al., 2011), qui est une collection de jeux de gènes dont l'annotation a été obtenue chez l'homme et la souris. Les bases de données humaines ont été largement plus enrichies en données fonctionnelles, notamment dues à une augmentation importante des études liées à ENCODE et aux développements technologiques associés (Hon and Carninci, 2020), que les bases dédiées aux animaux de rente. Afin de corriger une partie de ce biais, nous nous sommes affranchis des gènes non annotés dans le génome porcin en les éliminant des termes GO de la base de données MSigDB. Ce filtre a donc permis d'obtenir une base de données plus spécifique à notre espèce porcine malgré qu'elle soit alimentée par des travaux réalisés chez l'homme et la souris. Nous savons également que les connaissances accumulées ne sont pas équivalentes pour toutes les fonctions biologiques. Les voies métaboliques associées aux maladies humaines majeures dans nos sociétés comme les cancers ou la maladie d'Alzheimer par exemple, sont surreprésentées dans ces bases humaines et murines. Si pour certains caractères (et certaines fonctions), ces informations restent très intéressantes pour d'autres espèces comme le porc, l'étude de la CMJR correspond à une situation moins favorable que des caractères associés à la santé et la résistance immunitaire des animaux. Dans le cadre d'études de voies métaboliques enrichies en gènes de l'immunité chez le porc (Herrera-Uribe et al., 2020), les résultats obtenus avec GSEA et la base de données fonctionnelles humaines ont permis une caractérisation plus approfondie des systèmes biologiques et immunitaires du tissu sanguin porcin. Jack Dekkers et son doctorant m'ont également fait part de leur utilisation de GSEA avec une approche de recherche de voies métaboliques à partir de régions génomiques comme j'ai pu le développer, mais sur des données issues de prélèvement sanguin en lien à des maladies chez le porc et les résultats obtenus sont beaucoup plus riches et significatifs avec ce type de données. De ce fait, nous pouvons espérer qu'avec le développement croissant du consortium FAANG sur la centralisation et l'obtention de données fonctionnelles des espèces de rentes (Foissac et al., 2019), les bases de données fonctionnelles soient progressivement enrichies en annotation pour différents tissus et fonctions directement liées à des caractères étudiés chez nos espèces d'intérêt. Ainsi à l'avenir, ces études pourraient être plus sensibles si les outils développés tirent parti des bases de données fonctionnelles enrichies pour de nombreux organismes.

2. Limites de la combinaison des données génétiques et transcriptomiques

Dérive ou sélection ? : Les données génétiques utilisées dans mon projet sont issues du génotypage de plus de 1 600 animaux sélectionnés de manière divergente pour le caractère de la CMJR. Les régions génomiques présentant des fréquences alléliques divergentes entre les lignées ont été étudiées plus spécifiquement mais pour l'heure nous ne pouvons pas exclure que ces fortes évolutions de fréquences alléliques soient dues à de la dérive. La structure du dispositif de sélection basée sur l'utilisation de 6 reproducteurs mâles uniquement par génération et dans chaque lignée, est une situation particulièrement favorable à la dérive en raison de forts effets fondateurs. Dans ce travail nous n'avons pris en compte les évolutions de fréquences alléliques que dans des régions ciblées du génome : dans le chapitre 1 cette étude a porté uniquement sur les régions QTL, dans le chapitre 2 uniquement dans les régions où nous avions identifié des clusters de DEG. Une analyse préalable de recherche de traces de sélection sur l'ensemble du génome pourrait être réalisée. De nouvelles approches innovantes basées sur l'analyse de données temporelles (dans notre cas 2 X 10 générations) comme celle développée par Paris et al. (Paris et al., 2019) seraient intéressantes à appliquer sur ces données. Il n'est en effet pas exclu que d'autres régions d'intérêt existent (non détectées via les analyses GWAS, et/ou comprenant des clusters de DEG qui seraient exprimés dans d'autres tissus que ceux analysés).

Choix des tissus cibles : Les données transcriptomiques à notre disposition ont été obtenues sur 4 tissus : (*i*) le foie est un organe important dans le stockage des nutriments avec notamment son implication dans la mise en place des voies métaboliques liées au glucose et à son stockage en glycogène, et dans le métabolisme des lipides (Dodson et al., 2010; Nguyen et al., 2008), (*ii*) le muscle peut permettre la détection des voies métaboliques plus en lien avec la synthèse et l'accumulation des lipides, et ce type de processus biologiques a son importance pour la qualité de la viande (Li et al., 2012). Enfin, (*iii*) les tissus adipeux sont tout aussi intéressants à prendre en compte pour potentiellement identifier des voies métaboliques liées à la satiété (Yu and Kim, 2012). En effet le mécanisme de la satiété pourrait jouer un rôle important dans la caractérisation phénotypique des lignées divergentes. Ce type de mécanisme possède de nombreux régulateurs au niveau du cerveau, et obtenir des données transcriptomiques de cet organe serait tout à fait pertinent pour nos analyses. D'autre part, des tissus comme l'estomac, le pancréas ou l'intestin pourraient également avoir un rôle important dans la détection de voies métaboliques plus en lien avec les caractères de l'efficacité alimentaire, l'utilisation des nutriments et de leurs stockages, et la régulation de l'appétit (Yu and Kim, 2012). De ce point de vue-là, de prochaines données seront disponibles dans les mois à venir grâce au projet de Devailly et al. « ROSEpigs » qui souhaite étudier la satiété chez les porcs. Pour leurs études, 24 animaux (12 CMJR- et 12 CMJR+), issus du dispositif de sélection divergente pour la CMJR à la 11ème génération, ont été utilisés et des données de séquençage RNA-seq provenant de l'intestin (plus

précisément du duodénum, qui serait plus impliqué dans la régulation de la satiété à l'échelle du repas) vont donc être obtenues. Ainsi ces données complémentaires d'un point de vue des tissus et de la technique utilisée pour acquérir un transcriptome viendront alimenter les données fonctionnelles disponibles sur le dispositif expérimental et ainsi aider à mieux définir le caractère de l'efficacité alimentaire.

Combinaison de données génétiques et transcriptomiques : Les données transcriptomiques que nous avons utilisées proviennent du même dispositif expérimental, que les animaux utilisés pour les analyses GWAS. Les animaux possèdent donc bien le même fond génétique que les individus génotypés, mais aucun individu du dispositif GWAS ne possède de données transcriptomiques pour réaliser des études de eQTL. Ce type d'analyse pourrait être intéressante et puissante pour combiner les données fonctionnelles et la variabilité génétique des individus si la totalité des animaux génotypés possédait également des données transcriptomiques (Carmelo and Kadarmideen, 2020) et affiner notre compréhension de la biologie de l'efficacité alimentaire, comme le rapporte Higgins et al. (Higgins et al., 2018) chez les bovins. Ce type d'approche a toutefois un coût très élevé et uniquement dédié à l'acquisition de données transcriptomiques. En génétique humaine, plusieurs études ces derniers temps mettent en avant l'intérêt de diversifier la nature des données omiques pour affiner la caractérisation fonctionnelle et les voies métaboliques et protéiques, jouant un rôle majeur dans le phénotype d'intérêt. Plutôt que de mettre en place un dispositif dédié à des analyses eQTL, nous pourrions donc imaginer à l'avenir obtenir des données de génétique, du protéome, du métabolome et de l'épigénétique pour un lot réduit d'une cinquantaine d'individus. Ainsi le fait de connaître plus en détail les différents processus biologiques sous-jacents à un même groupe d'individus permettraient d'affiner nos connaissances sur la mise en place et l'expression du phénotype.

La difficulté porte alors sur la mise en commun de différents types de données. Dans la dernière étape de nos analyses réalisées avec GSEA, nous avons voulu ordonnancer la liste des gènes porcins sur la base de valeurs de p-value issues des analyses GWAS et de l'analyse transcriptomique. Par manque de temps, nous n'avons pas pu rechercher de solution dans la littérature et nous nous sommes donc limités à une première analyse brute. Cette dernière étude d'enrichissement n'a pas permis une bonne détection des voies métaboliques enrichies à la fois dans les approches de génétique et dans les approches transcriptomiques, car ce facteur "ordre dans la liste" des analyses GSEA impacte très fortement la liste des termes GO identifiés. N'ayant aucun critère pour ordonner les gènes les plus significatifs de l'analyse GWAS par rapport aux gènes les plus significatifs de l'analyse transcriptome, les résultats obtenus ne sont pas plus intéressants que ceux obtenus au préalables. Afin de mieux prendre en compte ces deux types de données, les approches d'intégration de données génétiques et transcriptomiques proposées par Keel et al. (Keel et al., 2020) ou Duarte (Duarte et al., 2019)

pourraient être testées afin de pondérer les marqueurs SNP utilisés pour les analyses GWAS en fonction des connaissances fonctionnelles.

Conclusion et perspectives

L'efficacité alimentaire est un caractère complexe qui a un impact économique direct sur la durabilité économique des élevages porcins en France et dans le monde. L'amélioration de l'efficacité alimentaire dans la production porcine et animale joue donc un rôle important pour la production d'une alimentation humaine durable dédiée à une population mondiale en constante augmentation. Afin d'élucider la question complexe de l'efficacité alimentaire, de nombreuses études ont été menées à INRAE et à l'ISU pour découvrir la base biologique, physiologique et génétique de la CMJR chez les porcs. Dans cette thèse, des lignées de porcs sélectionnées de manière divergente pour une amélioration de l'efficacité alimentaire (CMJR-) *versus* des animaux moins efficaces (CMJR+), ont été utilisées pour réaliser une cartographie fine de l'architecture du caractère de la CMJR. Les deux principaux objectifs de cette thèse étaient d'étudier le déterminisme génétique qui sous-tend l'efficacité alimentaire chez les porcs *via* des études d'association génotype-phénotype et d'explorer les fonctions associées aux gènes situés dans les régions en ségrégation pour ce caractère. Ce travail de thèse a également permis d'approfondir nos connaissances sur les régions génomiques associées à d'autres caractères phénotypiques d'intérêts corrélés à la CMJR. Les analyses GWAS réalisées et décrites dans le premier chapitre de ce manuscrit sont les premières études d'association réalisées à partir de l'ensemble du dispositif expérimental INRAE. Cette étude nous a ainsi permis de comparer ces lignées divergentes aux lignées ISU à l'échelle de leur génome et non uniquement de manière globale par la comparaison de paramètres de génétique quantitative. A l'issue de ce travail, d'autres analyses à partir du même dispositif pourraient être envisagées, afin d'affiner la précision de la cartographie des régions détectées ou l'interprétation de nos résultats.

Détection de traces de sélection : Au début de ce travail, une des questions posées était de savoir si dans les lignées INRAE, la variabilité de la CMJR résultait de la ségrégation des mêmes régions QTL ou si la pression de sélection exercée impactait des voies métaboliques différentes dans chacune des lignées. La comparaison des régions détectées *via* les analyses Global-GWAS, LRFI-GWAS et HRFI-GWAS nous ont amené à conclure qu'aucune région QTL pour la CMJR n'était partagée ; les régions QTL partagées étaient à chaque fois détectées pour des caractères différents. L'analyse des évolutions des fréquences alléliques des SNP des régions QTL semble indiquer qu'une part de cette différence résulterait d'un manque de puissance des marqueurs dans l'une ou l'autre des lignées, en raison d'une diminution progressive de l'informativité des marqueurs au cours de la sélection. Progressivement, la sélection d'allèles favorables pour le caractère induirait par effet d'entrainement dans les régions QTL de la CMJR une quasi fixation d'un des allèles au SNP (dans au moins une des deux lignées). Dans le génome, ce phénomène aboutit à "une trace de sélection". Dans des populations de faible taille

effective et sous sélection divergente, l'utilisation de méthodologies standards pour la détection des traces de sélection ne permet que rarement de conclure à un effet de la sélection en raison de l'effet majeur de la dérive sur les changements de fréquences des allèles. Au cours de sa thèse Emily Maunch, doctorante de l'ISU, a réalisé une mobilité internationale dans le laboratoire GenPhySE. Dans le cadre de cette mobilité, elle a utilisé les génotypes MD pour rechercher des traces de sélection au sein du dispositif INRAE et ISU en utilisant la méthode hapFLK (Fariello et al., 2013). Elle a ainsi identifié une région génomique significative et différente pour chaque dispositif sur le SSC2 dans les lignées ISU et sur le SSC13 dans les lignées INRAE. En combinant les deux dispositifs, les deux mêmes régions spécifiques aux deux dispositifs étaient détectées significatives. L'article détaillant ces analyses est en cours de préparation. Ainsi, compte tenu de la structure des lignées divergentes, ces analyses n'ont pas permis d'exclure la simple dérive comme source des évolutions de fréquences observées. Mais au cours de ma thèse de nouvelles méthodes ont récemment été développées qui pourraient être testées. Ces nouvelles approches sont basées sur l'exploitation de données génétiques (génotypes) temporelles (plusieurs échantillonnages au cours du temps) qui modélisent mieux la façon dont les génomes évoluent au sein d'une population. Cette méthode HMM développée par Paris et al. (Paris et al., 2019) serait plus adaptée à la détection de traces de sélection dans des petites populations et correspondrait donc à la suite des précédentes approches FLK (Bonhomme et al., 2010) et HapFLK (Fariello et al., 2013). Pour illustrer les performances statistiques de l'approche HMM sur un ensemble de données réelles, Paris et al. l'ont appliquée aux trajectoires de fréquences alléliques observées sur près de 40K SNP dans deux lignées de poulet sélectionnées de manière divergente pour le pH ultime intramusculaire. Cet ensemble de données a été préalablement analysé par Bihan-Duval et al. (Bihan-Duval et al., 2018) à l'aide des tests FLK (Bonhomme et al., 2010), afin de détecter un excès significatif de différenciation de fréquence des allèles et hapFLK pour la recherche des haplotypes entre les lignées divergentes (Fariello et al., 2013). Au final, cette approche novatrice de détection des traces de sélection a ainsi permis l'identification de 6 nouvelles régions en sélection, indiquant donc une plus grande précision de détection des traces de sélection dans une population présentant des données intermédiaires comme dans un dispositif de sélection divergente. Nous n'avons pas eu le temps de tester la méthode HMM sur nos données mais ce travail pourrait facilement être réalisé car il ne nécessite pas de données supplémentaires que celles que nous avons.

*Des résultats de GWAS avec des données de séque*nce : Les seconds travaux que nous pourrions entreprendre pour affiner nos résultats, seraient l'ajout de marqueurs supplémentaires. Les régions génomiques en ségrégation ont été étudiées dans le 1er chapitre avec des données génétiques HD d'environ 550K marqueurs. Grâce au projet ANR Micro-Feed, parmi les 32 animaux génotypés sur une puce HD, les génomes de 20 animaux ont également été séquencés à l'aide d'un séquenceur HiSeq.

Les animaux séquencés correspondent aux 12 mâles G0 (CMJR- et CMJR+) et à 8 femelles les plus contributrices des générations suivantes. Ces données génétiques haut débit pourraient donc être utilisées dans l'objectif de réaliser une imputation jusqu'à la séquence pour tous les individus génotypés. Ces génotypes imputés jusqu'à la séquence serviraient ensuite à l'obtention d'un génotype parental moyen pour les animaux réponses. Compte tenu des informations disponibles pour l'heure chez le porc, nous pourrions ainsi identifier +/- 5 millions de SNP par individu. Cette multiplication par 10 de la densité de marqueurs après imputation jusqu'à la séquence devrait être un fort atout pour la réalisation des études GWAS et la validation des régions génomiques déjà significatives avec les génotypes HD. Ces données augmenteront la probabilité de disposer de SNP en fort DL avec les mutations recherchées, voire de disposer des mutations elles-mêmes dans le jeu de marqueurs. Ainsi, la taille de la région génomique cible sera réduite et l'étude des gènes sous-jacents plus précise. De plus, nous pourrons également identifier de nouvelles régions en ségrégation dans des zones du génome qui seraient moins bien représentées sur les puces ADN et confirmer ou infirmer les QTL préalablement détectés. De plus, disposer de génomes spécifiques du dispositif entièrement séquencés présentent beaucoup d'avantages pour l'analyse des régions détectées. A l'issue de l'identification des régions QTL avec des données de séquence, une nouvelle recherche de gènes candidats présents aux positions des QTL pourra être menée. Si l'exploration des informations d'annotation seront réalisées à l'aide du génome de référence Duroc annoté et publié, la recherche de mutation sera réalisée dans les séquences des fondateurs des lignées. Nous pouvons ainsi imaginer détecter des CNV présents dans notre dispositif et non répertoriés jusqu'à présent ou explorer les conséquences fonctionnelles des variants identifiés à partir des séquences dans les régions QTL en utilisant l'outil « *Variant Effect Predictor* » (VEP) disponible sur Ensembl (McLaren et al., 2016). Récemment, plusieurs études basées sur des GWAS à partir de données de séquence ou des méta-analyses chez les bovins mettent en avant l'utilisation de cet outil dans l'analyse des QTL identifiés (Doyle et al., 2020; Marete et al., 2018; Pausch et al., 2017b). Enfin les données d'annotation du génome accumulées ou à venir issues des projets FAANG, permettant une meilleure caractérisation des éléments à différents niveaux biologiques, comme ceux agissant au niveau des protéines et de l'ARN, ou encore des éléments de régulation qui contrôlent les cellules, ou bien les circonstances dans lesquelles un gène est actif, pourraient également venir appuyer les résultats obtenus.

Des méta-analyses pour le caractère CMJR : La dernière piste qu'il serait intéressant d'explorer est l'augmentation de taille du dispositif. Aucune étude n'a encore été reportée à ce jour, quant à l'utilisation combinée de l'ensemble des données disponibles issues des multiples dispositifs publiés pour la CMJR chez les porcs. Dans un premier temps, une méta-analyse uniquement centrée sur les deux dispositifs de sélection divergente pour la CMJR de l'ISU et de l'INRAE pourrait être envisagée car

beaucoup de caractères phénotypiques sont communs entre les deux dispositifs et le Yorkshire et le Large-White sont deux races porcines apparentées. Nous pouvons espérer augmenter la puissance de détection pour certaines régions QTL identifiées pour l'heure que dans un seul dispositif et également confirmer et préciser la localisation de QTL déjà identifiés dans les deux dispositifs. Dans un second temps, une méta-analyse de tous les jeux de données de porcs pourrait être envisagée, même si différentes races porcines ont été étudiées dans les différents dispositifs. Il n'est pas exclu que certaines voies métaboliques soient communes quel que soit la race. Une telle approche a d'ores et déjà été menée en bovin par Duarte et al. (Duarte et al., 2019), qui ont alors pu trouver une voie significative (dégradation de la valine, de la leucine et de l'isoleucine) liée à la CMJR, comprenant 3 gènes (*MCCC1*, *AOX1* et *PCCA*) reportés dans trois études différentes. Les auteurs ont conclu qu'une méta-analyse basée sur les résultats de GWAS pourrait être une méthode appropriée pour découvrir des voies biologiques pour la CMJR en combinant différentes études.

Des études multi-espèces : Au-delà de la combinaison d'études réalisées sur le porc des données complémentaires pourraient être obtenues *via* une approche multi-espèces. Comme évoqué précédemment la CMJR est un caractère d'intérêt majeur pour toutes les filières animales et des travaux ont été publiés en génétique et en génomique pour plusieurs d'entre elles. A la fin du premier chapitre de cette thèse, nous avons réalisé une comparaison des régions génomiques obtenues dans notre étude à celles publiées en bovin et poulet. Peu de régions QTL sont partagées, mais cette approche pourrait être approfondie en ajoutant d'autres espèces comme le lapin voire l'homme et la souris. Même si certains processus biologiques pourraient être spécifiques à chacune des espèces cibles, nous pouvons imaginer que des voies métaboliques impliquées dans le stockage ou la mobilisation de l'énergie pourraient être communes à l'ensemble de ces espèces et contribuer à la variabilité de la CMJR. De plus, lors de nos comparaisons nous n'avons tenu compte que des résultats de GWAS, or nous pourrions également prendre en compte des travaux de transcriptomique et rechercher si les QTL identifiés chez différentes espèces pourraient partager des voies métaboliques communes.

Les pistes d'études complémentaires présentées jusqu'à présent sont destinées à poursuivre et améliorer la cartographie des régions QTL déterminant la variabilité génétique de la CMJR à partir de connaissances acquises sur le génome des individus. En complément, des données complémentaires sur la caractérisation du microbiote des individus pourrait être riche d'information au regard du caractère étudié.

Poursuivre la caractérisation du caractère CMJR :

Des données du microbiote intestinal : Ces dernières années un fort développement des méthodes d'intégration de données de microbiote dans diverses études a été mené et a notamment permis de mettre en évidence qu'un maximum d'énergie des rations alimentaires était capté par l'individu *via* son microbiote intestinal (Cardinelli et al., 2015). Goodrich et al. (Goodrich et al., 2014) ont suggéré chez l'homme que la susceptibilité héréditaire à l'obésité pourrait être en partie due à des facteurs génétiques contrôlant le microbiote. Knights et al. (Knights et al., 2014) ont ainsi identifié des polymorphismes de l'hôte corrélés avec la composition du microbiote intestinal lors d'études sur les maladies inflammatoires. Chez les poulets de chair, Mignon-Grasteau et al. (Mignon-Grasteau et al., 2015b) ont examiné un sous-ensemble de 144 poulets avec une efficacité alimentaire élevée et faible afin d'identifier les bactéries différenciant les groupes. Ils ont estimé des héritabilités modérées mais significatives pour certains microbiotes, et des corrélations génétiques significatives avec le caractère de digestibilité mesuré chez les animaux. En outre, ils ont identifié des régions génomiques du génome aviaire associées à l'abondance relative de certains microbiotes intestinaux. De plus, une étude de Garreau et al. (Garreau et al., 2019) a permis de mettre en évidence les différences pour la CMJR *via* la composition du microbiote intestinal des lapins. D'autres études menées par Ramayo-Caldas et al., Crespo-Piazuelo et al. et Camarinha et al. (Camarinha-Silva et al., 2017; Crespo-Piazuelo et al., 2019; Ramayo-Caldas et al., 2016) suggèrent un lien entre composition du microbiote intestinal et performance chez le porc en croissance, ainsi que sur sa santé. Chez les poulets, des études comparant de petits groupes de poulet et des individus présentant une efficacité alimentaire différente ont mis en évidence des compositions spécifiques de microbiote dans les fèces d'animaux plus efficaces (Singh et al., 2014).

Compte tenu de ces résultats, Gilbert et al. ont proposé l'analyse de données du microbiote intestinal *via* les fèces pour mieux caractériser le caractère CMJR pour l'efficacité alimentaire dans le projet ANR Micro-Feed. Les outils moléculaires haut-débit permettent aujourd'hui d'avoir accès à la composition du microbiote intestinal par l'amplification PCR puis le séquençage d'une portion du gène codant pour la sous-unité 16S de l'ARN ribosomique (ARNr 16S) variable entre bactérie. Pour cela les fèces des porcs en sélection de la génération 8 à la génération 10 ont été collectées en complément des données phénotypiques. Après séquençage l'ensemble des lectures sont comparées et assignées à des clusters de similarité (OTU : Operational Taxonomic Unit) et le nombre de lectures par cluster permet d'obtenir pour chaque échantillon une description de la population microbienne par la production d'une table d'abondance individuelle. A partir de ces prélèvements, l'objectif serait de rechercher les régions génomiques de l'hôte qui contrôleraient la composition du microbiote et/ou

l'efficacité alimentaire en combinant les profils OTU issus des tables d'abondance filtrées et normalisées et les génotypes des individus étudiés (Weissbrod et al., 2018).

Même si de multiples sources de données ont indiqué des interactions importantes entre l'hôte et le microbiome (Goodrich et al., 2014; Kurilshikov et al., 2017; Weissbrod et al., 2018), la part relative de ces interactions n'est pas claire, les études donnant des résultats quelque peu contrastés (Rothschild et al., 2018). Cela n'est peut-être pas surprenant étant donné la plasticité du microbiome face aux facteurs externes (maladies, conditions sanitaires). Dans ce contexte, une tâche essentielle et pourtant difficile, consiste à établir de véritables facteurs de causalité dans les associations observées entre l'environnement, la génétique de l'hôte et le microbiome lors de l'étude de caractères complexes. La combinaison de différents types de données du microbiome, c'est-à-dire protéomique, métabolomique et transcriptomique, pour compléter les données génomiques actuelles pourraient aider à éclairer ces interactions. Cependant, la combinaison de données complexes et de haut débit n'est pas simple, ce qui présente un autre défi (Awany et al., 2019).

Vers l'étude de nouveaux caractères

D'autres phénotypes disponibles : Dans le but de mieux comprendre les phénotypes de l'efficacité alimentaire chez les porcs, cette thèse était destinée à cartographier les QTL de la CMJR et de caractères corrélés à la CMJR. Néanmoins, d'autres données pourraient être intéressantes à analyser *via* des études GWAS comme l'ensemble des données de DAC récoltées pour chaque individu. Ces données demandent à être traitées avant leur possible utilisation mais elles permettraient de mieux étudier les comportements alimentaires des individus de chaque lignée, comme par exemple le nombre de passage au DAC et leur durée. En effet, la prise en compte de ces phénotypes en GWAS permettrait de caractériser plus précisément le comportement alimentaire des individus en fonction de leur lignée.

En dehors de la CMJR : Pour la filière porcine, les coûts d'alimentation représentent plus de la moitié des coûts de production totaux (chapitre 1). A cela s'ajoute des fluctuations importantes du coût des aliments pour animaux qui fragilisent la durabilité économique des élevages. Une des pistes d'amélioration de la durabilité des systèmes de production en filière porcine consisterait à remplacer le maïs et le soja dans les rations par des ingrédients moins coûteux. Toutefois les aliments de substitution sont souvent moins énergétiques et plus riches en fibres alimentaires, et peuvent réduire les performances des animaux qui les consomment. Ainsi en complément des travaux sur la CMJR, l'étude de l'efficacité digestive des animaux est un second paramètre qu'il serait nécessaire d'explorer ainsi que le lien entre ces deux paramètres. Mauch et al. (Mauch et al., 2018) ont étudié la digestibilité des porcs Yorkshire sélectionnés pour la CMJR et les auteurs ont conclu qu'une alimentation riche en

fibre réduirait la digestibilité, l'énergie brut, la matière sèche et l'azote, bien que les porcs CMJR- aient une digestibilité accrue de ces nutriments et de l'énergie par rapport aux porcs CMJR+ lorsqu'ils reçoivent cette alimentation. L'alimentation avec des aliments plus fibreux avait également réduit les réponses à la sélection pour les caractéristiques de performance. Ainsi, l'alimentation avec des aliments plus fibreux pourrait réduire le coût des intrants, mais pourrait ne pas être économiquement avantageux à long terme en raison de la réduction du taux de croissance, de l'augmentation de la consommation d'aliments et de la diminution des réactions aux caractéristiques de sélection ou de performance.

Depuis une dizaine de générations deux lignées divergentes sur la CMJR ont été conduites à INRAE. La majorité des études sur ces lignées ont été menées jusqu'à présent *via* des travaux de génétique quantitative et de physiologie afin de caractériser la réponse à la sélection et de comparer la physiologie des animaux CMJR- et CMJR+. Cette thèse rapporte les premiers travaux sur ces lignées de cartographie de QTL afin d'explorer l'architecture génétique du caractère. Beaucoup de travail reste à faire pour permettre l'identification des gènes sous-jacents à la variabilité. Mais les révolutions technologiques que l'on peut observer ces dernières années permettront d'accumuler de nouvelles informations qui alimenteront notre connaissance sur ce modèle. Nous avons tenté de combiner au cours de cette thèse des données génomiques et transcriptomiques ; la génétique systémique est également un domaine de recherche en plein essor et chez les animaux d'élevage elle pourra progressivement être implémentée grâce à l'acquisition de jeux de données multi-omiques combinées aux efforts de la communauté internationale sur l'amélioration des connaissances des génomes animaux, *via* des initiatives telles FAANG (http://www.faang.org). Cette thèse est un premier petit pas dans ce vaste domaine qui reste à explorer.

Références

Aerts, J., Megens, H.J., Veenendaal, T., Ovcharenko, I., Crooijmans, R., Gordon, L., Stubbs, L., and Groenen, M. (2007). Extent of linkage disequilibrium in chicken. CGR *117*, 338–345.

Aggrey, S.E., Karnuah, A.B., Sebastian, B., and Anthony, N.B. (2010). Genetic properties of feed efficiency parameters in meat-type chickens. Genetics Selection Evolution *42*, 25.

Agreste (2020). La consommation de viande en France en 2019. Agreste - Synthèse conjoncturelles.

Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S., and Lawlor, T.J. (2010). Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. Journal of Dairy Science *93*, 743–752.

Ali, B.M., Mey, Y. de, Bastiaansen, J.W.M., and Lansink, A.G.J.M.O. (2018). Effects of incorporating environmental cost and risk aversion on economic values of pig breeding goal traits. Journal of Animal Breeding and Genetics *135*, 194–207.

Aliakbari, A., Delpuech, E., Labrune, Y., Riquet, J., and Gilbert, H. (2020). The impact of training on data from genetically-related lines on the accuracy of genomic predictions for feed efficiency traits in pigs. Genetics Selection Evolution *52*, 57.

Amaral, A.J., Megens, H.-J., Crooijmans, R.P.M.A., Heuven, H.C.M., and Groenen, M.A.M. (2008). Linkage Disequilibrium Decay and Haplotype Block Structure in the Pig. Genetics *179*, 569–579.

Andersson, L., Archibald, A.L., Bottema, C.D., Brauning, R., Burgess, S.C., Burt, D.W., Casas, E., Cheng, H.H., Clarke, L., Couldrey, C., et al. (2015). Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. Genome Biology *16*, 57.

Archibald, A.L., Haley, C.S., Brown, J.F., Couperwhite, S., McQueen, H.A., Nicholson, D., Coppieters, W., Van de Weghe, A., Stratil, A., Winterø, A.K., et al. (1995). The PiGMaP consortium linkage map of the pig (Sus scrofa). Mammalian Genome *6*, 157–175.

Archibald, A.L., Bolund, L., Churcher, C., Fredholm, M., Groenen, M.A., Harlizius, B., Lee, K.-T., Milan, D., Rogers, J., Rothschild, M.F., et al. (2010). Pig genome sequence - analysis and publication strategy. BMC Genomics *11*, 438.

Ardlie, K.G., Kruglyak, L., and Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. Nature Reviews Genetics *3*, 299–309.

Arkfeld, E.K., Young, J.M., Johnson, R.C., Fedler, C.A., Prusa, K., Patience, J.F., Dekkers, J.C.M., Gabler, N.K., Lonergan, S.M., and Huff-Lonergan, E. (2015). Composition and quality characteristics of carcasses from pigs divergently selected for residual feed intake on high- or low-energy diets. Journal of Animal Science *93*, 2530–2545.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene Ontology: tool for the unification of biology. Nature Genetics *25*, 25–29.

Aunis, D., and Metz-Boutigue, M. (2001). Chromogranines : de la découverte à la fonction. Med Sci (Paris) *17*, 418.

Awany, D., Allali, I., Dalvie, S., Hemmings, S., Mwaikono, K.S., Thomford, N.E., Gomez, A., Mulder, N., and Chimusa, E.R. (2019). Host and Microbiome Genome-Wide Association Studies: Current State and Challenges. Front. Genet. *9*.

Azarpajouh, S., Colpoys, J., Dekkers, J., Gabler, N., Huff-Lonergan, E., Lonergan, S., Patience, J., and Johnson, A. (2017). How has selection for residual feed intake (RFI) affected nursery and finisher pig's feeding behavior and performance?

Badke, Y.M., Bates, R.O., Ernst, C.W., Fix, J., and Steibel, J.P. (2014). Accuracy of Estimation of Genomic Breeding Values in Pigs Using Low-Density Genotypes and Imputation. G3 Genes|Genomes|Genetics *4*, 623–631.

Bard, J.B.L., and Rhee, S.Y. (2004). Ontologies in biology: design, applications and future challenges. Nature Reviews Genetics *5*, 213–222.

Barea, R., Dubois, S., Gilbert, H., Sellier, P., van Milgen, J., and Noblet, J. (2010). Energy utilization in pigs selected for high and low residual feed intake. Journal of Animal Science *88*, 2062–2072.

Begli, H.E., Torshizi, R.V., Masoudi, A.A., Ehsani, A., and Jensen, J. (2016). Longitudinal analysis of body weight, feed intake and residual feed intake in F2 chickens. Livestock Science *184*, 28–34.

van den Berg, S., Vandenplas, J., van Eeuwijk, F.A., Bouwman, A.C., Lopes, M.S., and Veerkamp, R.F. (2019). Imputation to whole-genome sequence using multiple pig populations and its use in genome-wide association studies. Genetics Selection Evolution *51*, 2.

Bihan-Duval, E.L., Hennequet-Antier, C., Berri, C., Beauclercq, S.A., Bourin, M.C., Boulay, M., Demeure, O., and Boitard, S. (2018). Identification of genomic regions and candidate genes for chicken meat ultimate pH by combined detection of selection signatures and QTL. BMC Genomics *19*, 294.

van Binsbergen, R., Bink, M.C., Calus, M.P., van Eeuwijk, F.A., Hayes, B.J., Hulsegge, I., and Veerkamp, R.F. (2014). Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genetics Selection Evolution *46*, 41.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Stamatoyannopoulos, J.A., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799–816.

Boddicker, N., Gabler, N.K., Spurlock, M.E., Nettleton, D., and Dekkers, J.C.M. (2011). Effects of ad libitum and restricted feeding on early production performance and body composition of Yorkshire pigs selected for reduced residual feed intake. Animal *5*, 1344–1353.

Bodmer, W., and Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. Nature Genetics *40*, 695–701.

Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J., Blott, S., and SanCristobal, M. (2010). Detecting Selection in Population Trees: The Lewontin and Krakauer Test Extended. Genetics *186*, 241–262.

Browning, B.L., and Browning, S.R. (2009). A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. The American Journal of Human Genetics *84*, 210–223.

Bunter, K.L., Cai, W., Johnston, D.J., and Dekkers, J.C.M. (2010). Selection to reduce residual feed intake in pigs produces a correlated response in juvenile insulin-like growth factor-I concentration. Journal of Animal Science *88*, 1973–1981.

Burdick, J.T., Chen, W.-M., Abecasis, G.R., and Cheung, V.G. (2006). In silico method for inferring genotypes in pedigrees. Nature Genetics *38*, 1002–1004.

Cai, W., Casey, D.S., and Dekkers, J.C.M. (2008). Selection response and genetic parameters for residual feed intake in Yorkshire swine. Journal of Animal Science *86*, 287–298.

Camarinha-Silva, A., Maushammer, M., Wellmann, R., Vital, M., Preuss, S., and Bennewitz, J. (2017). Host Genome Influence on Gut Microbial Composition and Microbial Prediction of Complex Traits in Pigs. Genetics *206*, 1637–1644.

Cantor, R.M., Lange, K., and Sinsheimer, J.S. (2010). Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. The American Journal of Human Genetics *86*, 6–22.

Cardinelli, C.S., Sala, P.C., Alves, C.C., Torrinhas, R.S., and Waitzberg, D.L. (2015). Influence of Intestinal Microbiota on Body Weight Gain: a Narrative Review of the Literature. OBES SURG *25*, 346–353.

Carmelo, V.A., and Kadarmideen, H.N. (2020). Genetic variations (eQTLs) in muscle transcriptome and mitochondrial genes, and trans-eQTL molecular pathways in feed efficiency from Danish breeding pigs. BioRxiv 2020.04.17.047027.

Christensen, O.F., and Lund, M.S. (2010). Genomic prediction when some animals are not genotyped. Genetics Selection Evolution *42*, 2.

Christensen, O.F., Madsen, P., Nielsen, B., Ostersen, T., and Su, G. (2012). Single-step methods for genomic evaluation in pigs. Animal *6*, 1565–1571.

Civelek, M., and Lusis, A.J. (2014). Systems genetics approaches to understand complex traits. Nature Reviews Genetics *15*, 34–48.

Consortium, T.E.P. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. Science *306*, 636–640.

Crespo-Piazuelo, D., Migura-Garcia, L., Estellé, J., Criado-Mesas, L., Revilla, M., Castelló, A., Muñoz, M., García-Casco, J.M., Fernández, A.I., Ballester, M., et al. (2019). Association between the pig genome and its gut microbiota composition. Scientific Reports *9*, 8791.

Crews, D.H.D. (2005). Genetics of efficient feed utilization and national cattle evaluation: a review. Genet Mol Res *4*, 152–165.

Cruzen, S.M., Harris, A.J., Hollinger, K., Punt, R.M., Grubbs, J.K., Selsby, J.T., Dekkers, J.C.M., Gabler, N.K., Lonergan, S.M., and Huff-Lonergan, E. (2013). Evidence of decreased muscle protein turnover in gilts selected for low residual feed intake. Journal of Animal Science *91*, 4007–4016.

Daetwyler, H.D., Capitan, A., Pausch, H., Stothard, P., van Binsbergen, R., Brøndum, R.F., Liao, X., Djari, A., Rodriguez, S.C., Grohs, C., et al. (2014). Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nature Genetics *46*, 858–865.

Delanoue, E., and Roguet (2015). Acceptabilité sociale de l'élevage en France : recensement et analyse des principales controverses à partir des regards croisés de différents acteurs. INRAE Productions Animales *28*, 39–50.

Do, D.N., Strathe, A.B., Ostersen, T., Jensen, J., Mark, T., and Kadarmideen, H.N. (2013). Genome-Wide association study reveals genetic architecture of eating behavior in pigs and its implications for humans obesity by comparative mapping. PLOS ONE *8*, e71509.

Do, D.N., Ostersen, T., Strathe, A.B., Mark, T., Jensen, J., and Kadarmideen, H.N. (2014). Genome-wide association and systems genetic analyses of residual feed intake, daily feed consumption, backfat and weight gain in pigs. BMC Genetics *15*, 27.

Dodson, M.V., Hausman, G.J., Guan, L., Du, M., Rasmussen, T.P., Poulos, S.P., Mir, P., Bergen, W.G., Fernyhough, M.E., McFarland, D.C., et al. (2010). Lipid metabolism, adipocyte depot physiology and utilization of meat animals as experimental models for metabolic research. Int J Biol Sci *6*, 691–699.

Doss, S., Schadt, E.E., Drake, T.A., and Lusis, A.J. (2005). Cis-acting expression quantitative trait loci in mice. Genome Res. *15*, 681–691.

Doyle, J.L., Berry, D.P., Veerkamp, R.F., Carthy, T.R., Evans, R.D., Walsh, S.W., and Purfield, D.C. (2020). Genomic regions associated with muscularity in beef cattle differ in five contrasting cattle breeds. Genetics Selection Evolution *52*, 2.

Draghici, S., Khatri, P., Eklund, A.C., and Szallasi, Z. (2006). Reliability and reproducibility issues in DNA microarray measurements. Trends in Genetics *22*, 101–109.

Dronne, Y. (2018). Les matières premières agricoles pour l'alimentation humaine et animale : le monde. INRAE Productions Animales *31*, 165–180.

Druet, T., Schrooten, C., and de Roos, A.P.W. (2010). Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. Journal of Dairy Science *93*, 5443–5454.

Duarte, D. a. S., Newbold, C.J., Detmann, E., Silva, F.F., Freitas, P.H.F., Veroneze, R., and Duarte, M.S. (2019). Genome-wide association studies pathway-based meta-analysis for residual feed intake in beef cattle. Animal Genetics *50*, 150–153.

Duijvesteijn, N., Knol, E.F., Merks, J.W., Crooijmans, R.P., Groenen, M.A., Bovenhuis, H., and Harlizius, B. (2010). A genome-wide association study on androstenone levels in pigs reveals a cluster of candidate genes on chromosome 6. BMC Genetics *11*, 42.

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Ebana, Y., Sun, Y., Yang, X., Watanabe, T., Makita, S., Ozaki, K., Tanaka, T., Arai, H., and Furukawa, T. (2019). Pathway analysis with genome-wide association study (GWAS) data detected the association of atrial fibrillation with the mTOR signaling pathway. IJC Heart & Vasculature *24*, 100383.

Ellegren, H., Chowdhary, B.P., Johansson, M., Marklund, L., Fredholm, M., Gustavsson, I., and Andersson, L. (1994). A primary linkage map of the porcine genome reveals a low rate of genetic recombination. Genetics *137*, 1089–1100.

Excoffier, L., and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Molecular Biology and Evolution *12*, 921–927.

Fariello, M.I., Boitard, S., Naya, H., SanCristobal, M., and Servin, B. (2013). Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. Genetics *193*, 929–941.

Faure, J., Lefaucheur, L., Bonhomme, N., Ecolan, P., Meteau, K., Coustard, S.M., Kouba, M., Gilbert, H., and Lebret, B. (2013). Consequences of divergent selection for residual feed intake in pigs on muscle energy metabolism and meat quality. Meat Science *93*, 37–45.

Fehrmann, R.S.N., Jansen, R.C., Veldink, J.H., Westra, H.-J., Arends, D., Bonder, M.J., Fu, J., Deelen, P., Groen, H.J.M., Smolonska, A., et al. (2011). Trans-eQTLs Reveal That Independent Genetic Variants Associated with a Complex Phenotype Converge on Intermediate Genes, with a Major Role for the HLA. PLOS Genetics *7*, e1002197.

Feltus, F.A. (2014). Systems genetics: A paradigm to improve discovery of candidate genes and mechanisms underlying complex traits. Plant Science *223*, 45–48.

Foissac, S., Djebali, S., Munyard, K., Vialaneix, N., Rau, A., Muret, K., Esquerré, D., Zytnicki, M., Derrien, T., Bardou, P., et al. (2019). Multi-species annotation of transcriptome and chromatin structure in domesticated animals. BMC Biology *17*, 108.

FranceAgriMer (2020). FranceAgriMer : Consommation des produits carnés en 2019.

Freking, B.A., Murphy, S.K., Wylie, A.A., Rhodes, S.J., Keele, J.W., Leymaster, K.A., Jirtle, R.L., and Smith, T.P.L. (2002). Identification of the single base change causing the callipyge muscle hypertrophy phenotype, the only known example of polar overdominance in mammals. Genome Res. *12*, 1496–1506.

Garreau, H., Ruesche, J., Gilbert, H., Balmisse, E., Benitez, F., Richard, F., David, I., Drouilhet, L., and Zemb, O. (2019). Estimating direct genetic and maternal effects affecting rabbit growth and feed efficiency with a factorial design. Journal of Animal Breeding and Genetics *136*, 168–173.

The Gene Ontology Consortium (2019). The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Research *47*, D330–D338.

Gibbs, R.A., Belmont, J.W., Hardenbol, P., Willis, T.D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., Shen, Y., et al. (2003). The International HapMap Project. Nature *426*, 789–796.

Gilbert, H., Bidanel, J.-P., Gruand, J., Caritez, J.-C., Billon, Y., Guillouet, P., Lagant, H., Noblet, J., and Sellier, P. (2007). Genetic parameters for residual feed intake in growing pigs, with emphasis on genetic relationships with carcass and meat quality traits. J Anim Sci *85*, 3182–3188.

Gilbert, H., Guillouet, P., Noblet, J., van Milgen, J., and Brossard, L. (2009). Relations génétiques entre efficacité alimentaire et cinétiques de croissance et d'ingestion chez le porc Large White. Journées de la Recherche Porcine en France *41*, 1–6.

Gilbert, H., Billon, Y., Brossard, L., Faure, J., Gatellier, P., Gondret, F., Labussière, E., Lebret, B., Lefaucheur, L., Le Floch, N., et al. (2017a). Sélection pour la consommation alimentaire moyenne journalière résiduelle chez le porc : impacts sur les caractères et défis pour la filière. INRA Productions Animales 439–454.

Gilbert, H., Billon, Y., Brossard, L., Faure, J., Gatellier, P., Gondret, F., Labussière, E., Lebret, B., Lefaucheur, L., Floch, N.L., et al. (2017b). Review: divergent selection for residual feed intake in the growing pig. Animal *11*, 1427–1439.

GIS (2017). Efficience alimentaire des élevages.

Giuffra, E., and Tuggle, C.K. (2019). Functional Annotation of Animal Genomes (FAANG): Current Achievements and Roadmap. Annu. Rev. Anim. Biosci. *7*, 65–88.

Gondret, F., Vincent, A., Houée-Bigot, M., Siegel, A., Lagarrigue, S., Causeur, D., Gilbert, H., and Louveau, I. (2017). A transcriptome multi-tissue analysis identifies biological pathways and genes associated with variations in feed efficiency of growing pigs. BMC Genomics *18*, 244.

Goodrich, J.K., Waters, J.L., Poole, A.C., Sutter, J.L., Koren, O., Blekhman, R., Beaumont, M., Van Treuren, W., Knight, R., Bell, J.T., et al. (2014). Human Genetics Shape the Gut Microbiome. Cell *159*, 789–799.

Groenen, M. (2015). Development of a high-density Axiom® porcine genotyping array to meet research and commercial needs.

Groenen, M.A.M., Archibald, A.L., Uenishi, H., Tuggle, C.K., Takeuchi, Y., Rothschild, M.F., Rogel-Gaillard, C., Park, C., Milan, D., Megens, H.-J., et al. (2012). Analyses of pig genomes provide insight into porcine demography and evolution. Nature *491*, 393–398.

Grubbs, J.K., Fritchen, A.N., Huff-Lonergan, E., Dekkers, J.C.M., Gabler, N.K., and Lonergan, S.M. (2013). Divergent genetic selection for residual feed intake impacts mitochondria reactive oxygen species production in pigs1. Journal of Animal Science *91*, 2133–2140.

Gualdrón Duarte, J.L., Bates, R.O., Ernst, C.W., Raney, N.E., Cantet, R.J., and Steibel, J.P. (2013). Genotype imputation accuracy in a F2 pig population using high density and low density SNP panels. BMC Genetics *14*, 38.

Guo, X., Christensen, O.F., Ostersen, T., Wang, Y., Lund, M.S., and Su, G. (2016). Genomic prediction using models with dominance and imprinting effects for backfat thickness and average daily gain in Danish Duroc pigs. Genetics Selection Evolution *48*, 67.

Habier, D., Fernando, R.L., and Dekkers, J.C.M. (2009). Genomic Selection Using Low-Density Marker Panels. Genetics *182*, 343–353.

Hardie, L.C., VandeHaar, M.J., Tempelman, R.J., Weigel, K.A., Armentano, L.E., Wiggans, G.R., Veerkamp, R.F., de Haas, Y., Coffey, M.P., Connor, E.E., et al. (2017). The genetic and biological basis of feed efficiency in mid-lactation Holstein dairy cows. Journal of Dairy Science *100*, 9061–9075.

Harris, A.J., Patience, J.F., Lonergan, S.M., J.M. Dekkers, C., and Gabler, N.K. (2012). Improved nutrient digestibility and retention partially explains feed efficiency gains in pigs selected for low residual feed intake1. Journal of Animal Science *90*, 164–166.

Hayes, B.J., Bowman, P.J., Daetwyler, H.D., Kijas, J.W., and Werf, J.H.J. van der (2012). Accuracy of genotype imputation in sheep breeds. Animal Genetics *43*, 72–80.

Helgason, A., Yngvadóttir, B., Hrafnkelsson, B., Gulcher, J., and Stefánsson, K. (2005). An Icelandic example of the impact of population structure on association studies. Nature Genetics *37*, 90–95.

Herd, R.M., and Arthur, P.F. (2009). Physiological basis for residual feed intake. Journal of Animal Science *87*, E64–E71.

Herrera-Uribe, J., Byrne, K.A., Liu, H., Becker, S., Loving, C.L., and Tuggle, C.K. (2020). The transcriptional landscape of porcine peripheral blood immune cells. The Journal of Immunology *204*, 92.18-92.18.

Hess, C.W., Byerly, T.C., and Jull, M.A. (1941). The Efficiency of Feed Utilization by Barred Plymouth Rock and Crossbred Broilers. Poultry Science *20*, 210–216.

Hickey, J.M., Kinghorn, B.P., Tier, B., van der Werf, J.H., and Cleveland, M.A. (2012). A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genetics Selection Evolution *44*, 9.

Higgins, M.G., Fitzsimons, C., McClure, M.C., McKenna, C., Conroy, S., Kenny, D.A., McGee, M., Waters, S.M., and Morris, D.W. (2018). GWAS and eQTL analysis identifies a SNP associated with both residual feed intake and GFRA2 expression in beef cattle. Scientific Reports *8*, 14301.

Hill, W.G., and Robertson, A. (1966). The effect of linkage on limits to artificial selection. Genet Res *8*, 269–294.

Holden, M., Deng, S., Wojnowski, L., and Kulle, B. (2008). GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. Bioinformatics *24*, 2784–2785.

Hon, C.-C., and Carninci, P. (2020). Expanded ENCODE delivers invaluable genomic encyclopedia. Nature *583*, 685–686.

Hoque, M.A., and Suzuki, K. (2009). Genetics of residual feed intake in cattle and pigs: a review. Asian-Australasian Journal of Animal Sciences *22*, 747–755.

Hoque, M.A., Suzuki, K., Kadowaki, H., Shibata, T., and Oikawa, T. (2007). Genetic parameters for feed efficiency traits and their relationships with growth and carcass traits in Duroc pigs. Journal of Animal Breeding and Genetics *124*, 108–116.

Howie, B.N., Donnelly, P., and Marchini, J. (2009). A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLOS Genetics *5*, e1000529.

Hu, Z.-L., Park, C.A., and Reecy, J.M. (2019). Building a livestock genetic and genomic information knowledgebase through integrative developments of Animal QTLdb and CorrDB. Nucleic Acids Research *47*, D701–D710.

Huang, B.E., Reverter, A., Purvis, I., and Chapman, S. (2015). Meeting Report on the Challenge of Inference from Genome to Phenome. G3 Genes|Genomes|Genetics *5*, 1945–1947.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc *4*, 44–57.

Humphray, S.J., Scott, C.E., Clark, R., Marron, B., Bender, C., Camm, N., Davis, J., Jenks, A., Noon, A., Patel, M., et al. (2007). A high utility integrated map of the pig genome. Genome Biology *8*, R139.

IFIP (2011). Optimisation environnementale des élevages porcins de demain - Visions d'experts - La revue technique de l'IFIP.

IFIP (2020). IFIP : Coûts de revient internationaux en 2019, les résultats des élevages porcins en forte hausse.

Illumina (2015). PorcineSNP60 v2 Genotyping BeadChip : Description. Data Sheet: Agrigenomics.

International Human Genome Sequencing Consortium (2001). Erratum: Initial sequencing and analysis of the human genome. Nature *411*, 720–720.

Jansen, R.C. (2003). Studying complex biological systems using multifactorial perturbation. Nature Reviews Genetics *4*, 145–151.

Jansen, R.C., and Nap, J.-P. (2001). Genetical genomics: the added value from segregation. Trends in Genetics *17*, 388–391.

Jégou, M., Gondret, F., Vincent, A., Tréfeu, C., Gilbert, H., and Louveau, I. (2016). Whole blood transcriptomics is relevant to identify molecular changes in response to genetic selection for feed efficiency and nutritional status in the pig. PLOS ONE *11*, e0146550.

Jiao, S., Maltecca, C., Gray, K.A., and Cassady, J.P. (2014). Feed intake, average daily gain, feed efficiency, and real-time ultrasound traits in Duroc pigs: II. Genomewide association. J Anim Sci *92*, 2846–2860.

Jing, L., Hou, Y., Wu, H., Miao, Y., Li, X., Cao, J., Michael Brameld, J., Parr, T., and Zhao, S. (2015). Transcriptome analysis of mRNA and miRNA in skeletal muscle indicates an important network for differential Residual Feed Intake in pigs. Scientific Reports *5*, 11953.

Johnson, Z.B., Chewning, J.J., and Nugent, R.A., III (1999). Genetic parameters for production traits and measures of residual feed intake in large white swine. Journal of Animal Science *77*, 1679–1685.

Johnston, J., Kistemaker, G., and Sullivan, P.G. (2011). Comparison of different imputation methods. Interbull Bulletin *44*.

Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S., Freimer, N.B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nature Genetics *42*, 348–354.

Keel, B.N., Snelling, W.M., Lindholm-Perry, A.K., Oliver, W.T., Kuehn, L.A., and Rohrer, G.A. (2020). Using SNP weights derived from gene expression modules to improve GWAS power for feed efficiency in pigs. Front. Genet. *10*.

Kennedy, B.W., van der Werf, J.H., and Meuwissen, T.H. (1993). Genetic and statistical properties of residual feed intake. Journal of Animal Science *71*, 3239–3250.

Khansefid, M., Millen, C.A., Chen, Y., Pryce, J.E., Chamberlain, A.J., Vander Jagt, C.J., Gondro, C., and Goddard, M.E. (2017). Gene expression analysis of blood, liver, and muscle in cattle divergently selected for high and low residual feed intake1. Journal of Animal Science *95*, 4764–4775.

Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.-Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T., et al. (2005). Complement factor H polymorphism in age-related macular degeneration. Science *308*, 385–389.

Knights, D., Silverberg, M.S., Weersma, R.K., Gevers, D., Dijkstra, G., Huang, H., Tyler, A.D., van Sommeren, S., Imhann, F., Stempak, J.M., et al. (2014). Complex host genetics influence the microbiome in inflammatory bowel disease. Genome Med *6*, 107.

Koch, R.M., Swiger, L.A., Chambers, D., and Gregory, K.E. (1963). Efficiency of feed use in beef cattle. Journal of Animal Science *22*, 486–494.

Korte, A., and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. Plant Methods *9*, 29.

Kurilshikov, A., Wijmenga, C., Fu, J., and Zhernakova, A. (2017). Host Genetics and Gut Microbiome: Challenges and Perspectives. Trends in Immunology *38*, 633–647.

Labroue, F. (1995). Facteurs de variation génétiques de la prise alimentaire chez le porc en croissance : le point des connaissances. INRA production animale 12.

Labussière, E., Dubois, S., Gilbert, H., Thibault, J.N., Floc'h, N.L., Noblet, J., and Milgen, J. van (2015). Effect of inflammation stimulation on energy and nutrient utilization in piglets selected for low and high residual feed intake. Animal *9*, 1653–1661.

Lefaucheur, L., Lebret, B., Ecolan, P., Louveau, I., Damon, M., Prunier, A., Billon, Y., Sellier, P., and Gilbert, H. (2011). Muscle characteristics and meat quality traits are affected by divergent selection on residual feed intake in pigs1. Journal of Animal Science *89*, 996–1010.

Legarra, A., Aguilar, I., and Misztal, I. (2009). A relationship matrix including full pedigree and genomic information. Journal of Dairy Science *92*, 4656–4663.

Lewontin, R.C. (1964). The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. Genetics *49*, 49–67.

Li, W.Z., Zhao, S.M., Huang, Y., Yang, M.H., Pan, H.B., Zhang, X., Ge, C.R., and Gao, S.Z. (2012). Expression of lipogenic genes during porcine intramuscular preadipocyte differentiation. Research in Veterinary Science *93*, 1190–1194.

Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype Imputation. Annual Review of Genomics and Human Genetics *10*, 387–406.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics *27*, 1739–1740.

Lipkin, E., Mosig, M.O., Darvasi, A., Ezra, E., Shalom, A., Friedmann, A., and Soller, M. (1998). Quantitative trait locus mapping in dairy cattle by means of selective milk DNA pooling using dinucleotide microsatellite markers: analysis of milk protein percentage. Genetics *149*, 1557–1567.

Litt, M., and Luty, J.A. (1989). A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am J Hum Genet *44*, 397–401.

Liu, H., Nguyen, Y.T., Nettleton, D., Dekkers, J.C.M., and Tuggle, C.K. (2016). Post-weaning blood transcriptomic differences between Yorkshire pigs divergently selected for residual feed intake. BMC Genomics *17*, 73.

Lkhagvadorj, S., Qu, L., Cai, W., Couture, O.P., Barb, C.R., Hausman, G.J., Nettleton, D., Anderson, L.L., Dekkers, J.C.M., and Tuggle, C.K. (2009). Gene expression profiling of the short-term adaptive response

to acute caloric restriction in liver and adipose tissues of pigs differing in feed efficiency. American Journal of Physiology-Regulatory, Integrative and Comparative Physiology *298*, R494–R507.

Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Norton, H., et al. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. Nature Biotechnology *14*, 1675–1680.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The Genotype-Tissue Expression (GTEx) project. Nature Genetics *45*, 580–585.

Lu, Y., Vandehaar, M.J., Spurlock, D.M., Weigel, K.A., Armentano, L.E., Connor, E.E., Coffey, M., Veerkamp, R.F., de Haas, Y., Staples, C.R., et al. (2018). Genome-wide association analyses based on a multiple-trait approach for modeling feed efficiency. Journal of Dairy Science *101*, 3140–3154.

Ma, P., Brøndum, R.F., Zhang, Q., Lund, M.S., and Su, G. (2013). Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle. Journal of Dairy Science *96*, 4666–4677.

Malone, J.H., and Oliver, B. (2011). Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biol *9*, 34.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.

Marchini, J., and Howie, B. (2010). Genotype imputation for genome-wide association studies. Nature Reviews Genetics *11*, 499–511.

Marete, A.G., Guldbrandtsen, B., Lund, M.S., Fritz, S., Sahana, G., and Boichard, D. (2018). A meta-analysis including pre-selected sequence variants associated with seven traits in three french dairy cattle populations. Front. Genet. *9*.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. *18*, 1509–1517.

Mauch, E.D., Young, J.M., Serão, N.V.L., Hsu, W.L., Patience, J.F., Kerr, B.J., Weber, T.E., Gabler, N.K., and Dekkers, J.C.M. (2018). Effect of lower-energy, higher-fiber diets on pigs divergently selected for residual feed intake when fed higher-energy, lower-fiber diets1. Journal of Animal Science *96*, 1221–1236.

McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P.A., and Hirschhorn, J.N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature Reviews Genetics *9*, 356–369.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biology *17*, 122.

McRae, A.F., McEwan, J.C., Dodds, K.G., Wilson, T., Crawford, A.M., and Slate, J. (2002). Linkage disequilibrium in domestic sheep. Genetics *160*, 1113–1122.

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

Meunier-Salaün, M.C., Guérin, C., Billon, Y., Sellier, P., Noblet, J., and Gilbert, H. (2014). Divergent selection for residual feed intake in group-housed growing pigs: characteristics of physical and behavioural activity according to line and sex. Animal *8*, 1898–1906.

Meuwissen, T.H.E., Hayes, B.J., and Goddard, M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics *157*, 1819–1829.

Mignon-Grasteau, S., Rideau, N., Gabriel, I., Chantry-Darmon, C., Boscher, M.-Y., Sellier, N., Chabault, M., Le Bihan-Duval, E., and Narcy, A. (2015a). Detection of QTL controlling feed efficiency and excretion in chickens fed a wheat-based diet. Genet Sel Evol *47*, 74.

Mignon-Grasteau, S., Narcy, A., Rideau, N., Chantry-Darmon, C., Boscher, M.-Y., Sellier, N., Chabault, M., Konsak-Ilievski, B., Bihan-Duval, E.L., and Gabriel, I. (2015b). Impact of selection for digestive efficiency on microbiota composition in the chicken. PLOS ONE *10*, e0135488.

Montefiori, L.E., Sobreira, D.R., Sakabe, N.J., Aneas, I., Joslin, A.C., Hansen, G.T., Bozek, G., Moskowitz, I.P., McNally, E.M., and Nóbrega, M.A. (2018). A promoter interaction map for cardiovascular disease genetics. ELife *7*, e35788.

Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., Kaul, R., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature *583*, 699–710.

Mootha, V.K., Lindgren, C.M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., et al. (2003). PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genetics *34*, 267–273.

Mrode, R.A., and Kennedy, B.W. (1993). Genetic variation in measures of food efficiency in pigs and their genetic relationships with growth rate and backfat. Animal Science *56*, 225–232.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science *320*, 1344–1349.

Nguyen, N.H., McPhee, C.P., and Wade, C.M. (2005). Responses in residual feed intake in lines of Large White pigs selected for growth rate on restricted feeding (measured on ad libitum individual feeding). Journal of Animal Breeding and Genetics *122*, 264–270.

Nguyen, P., Leray, V., Diez, M., Serisier, S., Bloc'h, J.L., Siliart, B., and Dumon, H. (2008). Liver lipid metabolism. Journal of Animal Physiology and Animal Nutrition *92*, 272–283.

Nica, A.C., Montgomery, S.B., Dimas, A.S., Stranger, B.E., Beazley, C., Barroso, I., and Dermitzakis, E.T. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLOS Genetics *6*, e1000895.

Nsengimana, J., Baret, P., Haley, C.S., and Visscher, P.M. (2004). Linkage disequilibrium in the domesticated pig. Genetics *166*, 1395–1404.

OCDE and Organisation des Nations Unies pour l'alimentation et l'agriculture (2020). Perspectives agricoles de l'OCDE et de la FAO 2020-2029 (OCDE).

OCDE-FAO (2019). Chapitre 6 Viande.

Onteru, S.K., Gorbach, D.M., Young, J.M., Garrick, D.J., Dekkers, J.C.M., and Rothschild, M.F. (2013). Whole genome association studies of residual feed intake and related traits in the pig. PLOS ONE *8*, e61756.

Ostersen, T., Christensen, O.F., Henryon, M., Nielsen, B., Su, G., and Madsen, P. (2011). Deregressed EBV as the response variable yield more reliable genomic predictions than traditional EBV in pure-bred pigs. Genet Sel Evol *43*, 38.

Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T., Sato, H., Sato, H., Hori, M., Nakamura, Y., et al. (2002). Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. Nature Genetics *32*, 650–654.

Pakdel, A., Arendonk, J.A.M.V., Vereijken, A.L.J., and Bovenhuis, H. (2005). Genetic parameters of ascites-related traits in broilers: correlations with feed efficiency and carcase traits. British Poultry Science *46*, 43–53.

Paris, C., Servin, B., and Boitard, S. (2019). Inference of Selection from Genetic Time Series Using Various Parametric Approximations to the Wright-Fisher Model. G3: Genes, Genomes, Genetics *9*, 4073–4086.

Pausch, H., MacLeod, I.M., Fries, R., Emmerling, R., Bowman, P.J., Daetwyler, H.D., and Goddard, M.E. (2017a). Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle. Genetics Selection Evolution *49*, 24.

Pausch, H., Emmerling, R., Gredler-Grandl, B., Fries, R., Daetwyler, H.D., and Goddard, M.E. (2017b). Meta-analysis of sequence-based association studies across three cattle breeds reveals 25 QTL for fat and protein percentages in milk at nucleotide resolution. BMC Genomics *18*, 853.

Pazin, M.J. (2015). Using the ENCODE Resource for Functional Annotation of Genetic Variants. Cold Spring Harb Protoc *2015*.

Phocas, F., Agabriel, J., Dupont-Nivet, M., Geurden, J., Médale, F., MIGNON-GRASTEAU, S., GILBERT, H., and DOURMAD, J.Y. (2014). Le phénotypage de l'efficacité alimentaire et de ses composantes, une nécessité pour accroître l'efficience des productions animales. INRAE Productions Animales *27*, 235–248.

Ponsuksili, S., Jonas, E., Murani, E., Phatsara, C., Srikanchai, T., Walz, C., Schwerin, M., Schellander, K., and Wimmers, K. (2008). Trait correlated expression combined with expression QTL analysis reveals biological pathways and candidate genes affecting water holding capacity of muscle. BMC Genomics *9*, 367.

Ponsuksili, S., Murani, E., Schwerin, M., Schellander, K., and Wimmers, K. (2010). Identification of expression QTL (eQTL) of genes expressed in porcine M. longissimus dorsi and associated with meat quality traits. BMC Genomics *11*, 572.

Ponsuksili, S., Murani, E., Trakooljul, N., Schwerin, M., and Wimmers, K. (2014). Discovery of candidate genes for muscle traits based on GWAS supported by eQTL-analysis. Int J Biol Sci *10*, 327–337.

Ponsuksili, S., Zebunke, M., Murani, E., Trakooljul, N., Krieter, J., Puppe, B., Schwerin, M., and Wimmers, K. (2015). Integrated Genome-wide association and hypothalamus eQTL studies indicate a link between the circadian rhythm-related gene PER1 and coping behavior. Scientific Reports *5*, 16264.

Porto-Neto, L.R., Kijas, J.W., and Reverter, A. (2014). The extent of linkage disequilibrium in beef cattle breeds using high-density SNP genotypes. Genetics Selection Evolution *46*, 22.

Poux, S., and Gaudet, P. (2017). Best practices in manual annotation with the gene ontology. In The Gene Ontology Handbook, C. Dessimoz, and N. Škunca, eds. (New York, NY: Springer), pp. 41–54.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics *81*, 559–575.

Ramayo-Caldas, Y., Mach, N., Lepage, P., Levenez, F., Denis, C., Lemonnier, G., Leplat, J.-J., Billon, Y., Berri, M., Doré, J., et al. (2016). Phylogenetic network analysis applied to pig gut microbiota identifies an ecosystem structure linked with growth traits. The ISME Journal *10*, 2973–2977.

Ramos, A.M., Crooijmans, R.P.M.A., Affara, N.A., Amaral, A.J., Archibald, A.L., Beever, J.E., Bendixen, C., Churcher, C., Clark, R., Dehais, P., et al. (2009). Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. PLOS ONE *4*, e6524.

Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R., et al. (2001). Linkage disequilibrium in the human genome. Nature *411*, 199–204.

Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., Wadi, L., Meyer, M., Wong, J., Xu, C., et al. (2019). Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nature Protocols *14*, 482–517.

Ritchie, H., and Roser, M. (2017). Meat and dairy production. Our World in Data.

Roguet, C., Duflot, B., and Rieu, M. (2017). Évolution des modèles d'élevage de porcs en Europe et impacts sur les performances technico-économiques. Économie rurale. Agricultures, alimentations, territoires 73–86.

Rohrer, G.A., Alexander, L.J., Keele, J.W., Smith, T.P., and Beattie, C.W. (1994). A microsatellite linkage map of the porcine genome. Genetics *136*, 231–245.

Rothschild, D., Weissbrod, O., Barkan, E., Kurilshikov, A., Korem, T., Zeevi, D., Costea, P.I., Godneva, A., Kalka, I.N., Bar, N., et al. (2018). Environment dominates over host genetics in shaping human gut microbiota. Nature *555*, 210–215.

Sadler, L.J., Johnson, A.K., Lonergan, S.M., Nettleton, D., and Dekkers, J.C.M. (2011). The effect of selection for residual feed intake on general behavioral activity and the occurrence of lesions in Yorkshire gilts1. Journal of Animal Science *89*, 258–266.

Saintilan, R., Merour, I., Milgen, J.V., and Gilbert, H. (2012). Sélection pour l'efficacité alimentaire chez le porc en croissance : opportunités et conséquences de l'utilisation de la consommation moyenne journalière résiduelle dans les populations en sélection collective. *44*, 13–18.

Sargolzaei, M., Chesnais, J.P., and Schenkel, F.S. (2014). A new approach for efficient genotype imputation using information from relatives. BMC Genomics *15*, 478.

Schadt, E.E., Lamb, J., Yang, X., Zhu, J., Edwards, S., GuhaThakurta, D., Sieberts, S.K., Monks, S., Reitman, M., Zhang, C., et al. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. Nature Genetics *37*, 710–717.

Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S., and Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. Genome Res. *22*, 1748–1759.

Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science *270*, 467–470.

Schneider, J.F., Rempel, L.A., Snelling, W.M., Wiedmann, R.T., Nonneman, D.J., and Rohrer, G.A. (2012). Genome-wide association study of swine farrowing traits. Part II: Bayesian analysis of marker data1,2. Journal of Animal Science *90*, 3360–3367.

Schook, L.B., Beever, J.E., Rogers, J., Humphray, S., Archibald, A., Chardon, P., Milan, D., Rohrer, G., and Eversole, K. (2005). Swine Genome Sequencing Consortium (SGSC): a strategic roadmap for sequencing the pig genome. Comparative and Functional Genomics *6*, 251–255.

Schroyen, M., and Tuggle, C.K. (2015). Current transcriptomics in pig immunity research. Mamm Genome *26*, 1–20.

Seabury, C.M., Oldeschulte, D.L., Saatchi, M., Beever, J.E., Decker, J.E., Halley, Y.A., Bhattarai, E.K., Molaei, M., Freetly, H.C., Hansen, S.L., et al. (2017). Genome-wide association study for feed efficiency and growth traits in U.S. beef cattle. BMC Genomics *18*, 386.

Shin, D., Chang, S.Y., Bogere, P., Won, K., Choi, J.-Y., Choi, Y.-J., Lee, H.K., Hur, J., Park, B.-Y., Kim, Y., et al. (2019). Beneficial roles of probiotics on the modulation of gut microbiota and immune response in pigs. PLoS One *14*.

van der Sijde, M.R., Ng, A., and Fu, J. (2014). Systems genetics: From GWAS to disease pathways. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease *1842*, 1903–1909.

Singh, K.M., Shah, T.M., Reddy, B., Deshpande, S., Rank, D.N., and Joshi, C.G. (2014). Taxonomic and gene-centric metagenomics of the fecal microbiome of low and high feed conversion ratio (FCR) broilers. J Appl Genetics *55*, 145–154.

Smith, R.M., Gabler, N.K., Young, J.M., Cai, W., Boddicker, N.J., Anderson, M.J., Huff-Lonergan, E., Dekkers, J.C.M., and Lonergan, S.M. (2011). Effects of selection for decreased residual feed intake on composition and quality of fresh pork1. Journal of Animal Science *89*, 192–200.

Snyder, M.P., Gingeras, T.R., Moore, J.E., Weng, Z., Gerstein, M.B., Ren, B., Hardison, R.C., Stamatoyannopoulos, J.A., Graveley, B.R., Feingold, E.A., et al. (2020). Perspectives on ENCODE. Nature *583*, 693–698.

Soleimani, T., and Gilbert, H. (2020). Evaluating environmental impacts of selection for residual feed intake in pigs. Animal *14*, 2598–2608.

Song, H., Ye, S., Jiang, Y., Zhang, Z., Zhang, Q., and Ding, X. (2019). Using imputation-based whole-genome sequencing data to improve the accuracy of genomic prediction for combined populations in pigs. Genetics Selection Evolution *51*, 58.

Speakman, J.R., Loos, R.J.F., O'Rahilly, S., Hirschhorn, J.N., and Allison, D.B. (2018). GWAS for BMI: a treasure trove of fundamental insights into the genetic basis of obesity. International Journal of Obesity *42*, 1524–1531.

Spencer, C.C.A., Su, Z., Donnelly, P., and Marchini, J. (2009). Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLOS Genetics *5*, e1000477.

Sridhar, G.R. (2018). Encode, Decode and Diabetes. In Cognitive Science and Health Bioinformatics: Advances and Applications, R.B. Korrapati, Ch. Divakar, and G.L. Devi, eds. (Singapore: Springer), pp. 47–55.

Stephens, M., Smith, N.J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. The American Journal of Human Genetics *68*, 978–989.

Stunnenberg, H.G., Abrignani, S., Adams, D., de Almeida, M., Altucci, L., Amin, V., Amit, I., Antonarakis, S.E., Aparicio, S., Arima, T., et al. (2016). The International Human Epigenome Consortium: a blueprint for scientific collaboration and discovery. Cell *167*, 1145–1149.

Su, A.I., Cooke, M.P., Ching, K.A., Hakak, Y., Walker, J.R., Wiltshire, T., Orth, A.P., Vega, R.G., Sapinoso, L.M., Moqrich, A., et al. (2002). Large-scale analysis of the human and mouse transcriptomes. PNAS *99*, 4465–4470.

Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., et al. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. PNAS *101*, 6062–6067.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences *102*, 15545–15550.

Sun, Y.V., and Hu, Y.-J. (2016). Chapter Three - Integrative Analysis of Multi-omics Data for Discovery and Functional Studies of Complex Human Diseases. In Advances in Genetics, T. Friedmann, J.C. Dunlap, and S.F. Goodwin, eds. (Academic Press), pp. 147–190.

Suravajhala, P., Kogelman, L.J.A., and Kadarmideen, H.N. (2016). Multi-omic data integration and analysis using systems genomics approaches: methods and applications in animal production, health and welfare. Genetics Selection Evolution *48*, 38.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. Nature Reviews Genetics *20*, 467–484.

Taşan, M., Musso, G., Hao, T., Vidal, M., MacRae, C.A., and Roth, F.P. (2015). Selecting causal genes from genome-wide association studies via functionally coherent subnetworks. Nature Methods *12*, 154–159.

Thorisson, G.A., Smith, A.V., Krishnan, L., and Stein, L.D. (2005). The International HapMap Project Web site. Genome Res. *15*, 1592–1593.

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn) *2015*, 68–77.

Torkamani, A., Topol, E.J., and Schork, N.J. (2008). Pathway analysis of seven common diseases assessed by genome-wide association. Genomics *92*, 265–272.

Valicherla, G.R., Hossain, Z., Mahata, S.K., and Gayen, J.R. (2013). Pancreastatin is an endogenous peptide that regulates glucose homeostasis. Physiological Genomics *45*, 1060–1071.

Van Laere, A.-S., Nguyen, M., Braunschweig, M., Nezer, C., Collette, C., Moreau, L., Archibald, A.L., Haley, C.S., Buys, N., Tally, M., et al. (2003). A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. Nature *425*, 832–836.

VanRaden, Paul M., O'Connell, Jeffrey R., Wiggans, G.R., and Weigel, K.A. (2011). Genomic evaluations with many more genotypes. Genetics Selection Evolution *43*, 10.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The Sequence of the Human Genome. Science *291*, 1304–1351.

Villar, D., Frost, S., Deloukas, P., and Tinker, A. (2020). The contribution of non-coding regulatory elements to cardiovascular disease. Open Biology *10*, 200088.

Vincent, A., Louveau, I., Gondret, F., Tréfeu, C., Gilbert, H., and Lefaucheur, L. (2015). Divergent selection for residual feed intake affects the transcriptomic and proteomic profiles of pig skeletal muscle12. Journal of Animal Science *93*, 2745–2758.

Von Felde, A., Roehe, R., Looft, H., and Kalm, E. (1996). Genetic association between feed intake and feed intake behaviour at different stages of growth of group-housed boars. Livestock Production Science *47*, 11–22.

Wall, E., Simm, G., and Moran, D. (2010). Developing breeding schemes to assist mitigation of greenhouse gas emissions. Animal *4*, 366–376.

Walley, A.J., Jacobson, P., Falchi, M., Bottolo, L., Andersson, J.C., Petretto, E., Bonnefond, A., Vaillant, E., Lecoeur, C., Vatin, V., et al. (2012). Differential coexpression analysis of obesity-associated networks in human subcutaneous adipose tissue. International Journal of Obesity *36*, 137–147.

Wang, K., Zhang, H., Kugathasan, S., Annese, V., Bradfield, J.P., Russell, R.K., Sleiman, P.M.A., Imielinski, M., Glessner, J., Hou, C., et al. (2009a). Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn disease. The American Journal of Human Genetics *84*, 399–405.

Wang, Z., Gerstein, M., and Snyder, M. (2009b). RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics *10*, 57–63.

Warr, A., Affara, N., Aken, B., Beiki, H., Bickhart, D.M., Billis, K., Chow, W., Eory, L., Finlayson, H.A., Flicek, P., et al. (2020). An improved pig reference genome sequence to enable pig genetics and genomics research. GigaScience *9*, 1–14.

Watanabe, K., Kuramitsu, S., Posey, A.D.J., and June, C.H. (2018). Expanding the Therapeutic Window for CAR T Cell Therapy in Solid Tumors: The Knowns and Unknowns of CAR T Cell Biology. Front. Immunol. *9*.

Weigel, K.A., de los Campos, G., Vazquez, A.I., Rosa, G.J.M., Gianola, D., and Van Tassell, C.P. (2010). Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. Journal of Dairy Science *93*, 5423–5435.

Weiss, K.M., and Clark, A.G. (2002). Linkage disequilibrium and the mapping of complex human traits. Trends in Genetics *18*, 19–24.

Weissbrod, O., Rothschild, D., Barkan, E., and Segal, E. (2018). Host genetics and microbiome associations through the lens of genome wide association studies. Current Opinion in Microbiology *44*, 9–19.

Wu, A.R., Neff, N.F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M.E., Mburu, F.M., Mantalas, G.L., Sim, S., Clarke, M.F., et al. (2014). Quantitative assessment of single-cell RNA-sequencing methods. Nature Methods *11*, 41–46.

Yang, E.-W., Bahn, J.H., Hsiao, E.Y.-H., Tan, B.X., Sun, Y., Fu, T., Zhou, B., Van Nostrand, E.L., Pratt, G.A., Freese, P., et al. (2019). Allele-specific binding of RNA-binding proteins reveals functional genetic variants in the RNA. Nature Communications *10*, 1338.

Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. The American Journal of Human Genetics *88*, 76–82.

Yon Rhee, S., Wood, V., Dolinski, K., and Draghici, S. (2008). Use and misuse of the gene ontology annotations. Nature Reviews Genetics *9*, 509–515.

Young, J.M., Cai, W., and Dekkers, J.C.M. (2011). Effect of selection for residual feed intake on feeding behavior and daily feed intake patterns in Yorkshire swine1. Journal of Animal Science *89*, 639–647.

Yu, J.H., and Kim, M.-S. (2012). Molecular mechanisms of appetite regulation. Diabetes Metab J *36*, 391–398.

Yuan, J., Dou, T., Ma, M., Yi, G., Chen, S., Qu, L., Shen, M., Qu, L., Wang, K., and Yang, N. (2015a). Genetic parameters of feed efficiency traits in laying period of chickens. Poultry Science *94*, 1470–1475.

Yuan, J., Wang, K., Yi, G., Ma, M., Dou, T., Sun, C., Qu, L.-J., Shen, M., Qu, L., and Yang, N. (2015b). Genome-wide association studies for feed intake and efficiency in two laying periods of chickens. Genetics Selection Evolution *47*, 82.

Zhang, F., and Drabier, R. (2012). IPAD: the Integrated Pathway Analysis Database for systematic enrichment analysis. BMC Bioinformatics *13*, S7.

Zhang, C., Kemp, R.A., Stothard, P., Wang, Z., Boddicker, N., Krivushin, K., Dekkers, J., and Plastow, G. (2018). Genomic evaluation of feed efficiency component traits in Duroc pigs using 80K, 650K and whole-genome sequence variants. Genetics Selection Evolution *50*, 14.

Zhang, J., Lee, D., Dhiman, V., Jiang, P., Xu, J., McGillivray, P., Yang, H., Liu, J., Meyerson, W., Clarke, D., et al. (2020). An integrative ENCODE resource for cancer genomics. Nature Communications *11*, 3696.

Zhang, K., Chang, S., Guo, L., and Wang, J. (2015). I-GSEA4GWAS v2: a web server for functional analysis of SNPs in trait-associated pathways identified from genome-wide association study. Protein Cell *6*, 221–224.

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nat Genet *44*, 821–824.

Annexe 1

Trait	Type_Analyse	Intra-Analysis_Region	QTL_Region(trait x region)	SNP_QTL	SSC	Position_SNP_QTL	-log10(p-value)	QTL_effect(Beta)
carcBFT	Global-GWAS	8_9	8_9	AX-116099529	1	8130617	4.65	1.022
RFI	LRFI-GWAS	62_64	62_64	INRA0002667	1	62781138	4.52	0.049
RFI	LRFI-GWAS	65_66	65_66	DRGA0001106	1	65460886	4.58	0.05
RFI	LRFI-GWAS	76_77	76_77	AX-116117019	1	76828100	4.58	-0.043
a*_GM	Global-GWAS	86_87	86_87	AX-116119439	1	86451439	4.68	-1.152
a*_GM	Global-GWAS	92_93	92_93	AX-116633511	1	92771467	5.37	-1.38
DFI	Global-GWAS	147_148	147_151	AX-116130220	1	147312638	4.56	0.057
DFI	LRFI-GWAS	147_151	147_151	AX-116130853	1	149982825	4.85	-0.062
DFI	LRFI-GWAS	155_157	155_157	AX-116132180	1	155861537	4.57	-0.06
DFI	HRFI-GWAS	182_183	182_183	AX-116137586	1	182704229	4.71	0.092
DFI	HRFI-GWAS	185_187	185_187	AX-116138339	1	185715484	4.51	0.089
carcBFT	LRFI-GWAS	193_194	193_194	AX-116139873	1	193721759	4.59	-1.381
carcBFT	Global-GWAS	196_197	196_197	AX-116140272	1	196631045	5.11	-0.963
carcBFT	Global-GWAS	199_201	199_201	AX-116704657	1	199116470	4.66	-0.925
LMCcalc	HRFI-GWAS	222_223	222_223	AX-116146296	1	222498254	4.58	1.166
BF_W	HRFI-GWAS	222_223	222_223	AX-116146347	1	222668686	5.11	-0.21
WHC	HRFI-GWAS	249_250	249_250	AX-116152603	1	249584786	4.52	17.31
ADG	Global-GWAS	254_255	254_255	AX-116153742	1	254516626	4.63	-0.021
WHC	LRFI-GWAS	266_267	266_267	WU_10.2_1_300796817	1	266942535	4.52	-13.604
L*_GS	Global-GWAS	268_269	268_269	AX-116639159	1	268214672	4.78	-1.63
MQI	Global-GWAS	268_269	268_269	AX-116156798	1	268317623	5.13	1.061
Ham_W	HRFI-GWAS	293_294	293_294	AX-116162950	2	18002476	4.9	0.197
L*_GS	LRFI-GWAS	300_301	300_301	AX-116164920	2	25774144	4.63	1.13
Ham_W	Global-GWAS	306_308	306_308	AX-116166602	2	32914185	5.05	-0.181
L*_GS	LRFI-GWAS	307_308	306_308	AX-116166609	2	32953432	6.14	1.582
Ham_W	Global-GWAS	309_310	309_310	AX-116166897	2	34113988	4.56	-0.136
L*_GS	LRFI-GWAS	309_310	309_310	AX-116167092	2	34827646	4.51	-1.048
Ham_W	Global-GWAS	313_315	313_316	WU_10.2_2_41888756	2	39017255	5.14	-0.171
a*_GS	LRFI-GWAS	315_316	313_316	AX-116168477	2	40524569	4.57	-0.562
BF_W	Global-GWAS	349_350	349_350	AX-116173979	2	74523253	4.55	0.127
BF_W	Global-GWAS	352_353	352_353	AX-116712862	2	77405635	4.56	-0.127
BF_W	Global-GWAS	354_357	354_359	WU_10.2_2_83165289_80k	2	81513887	5.01	-0.134
BF_W	LRFI-GWAS	356_359	354_359	WU_10.2_2_83165289_80k	2	81513887	5.87	-0.179
LMCcalc	Global-GWAS	354_357	354_359	WU_10.2_2_83165289_80k	2	81513887	5.06	0.89
BF_W	Global-GWAS	358_359	354_359	AX-116655949	2	83558137	5.34	-0.14
LMCcalc	Global-GWAS	358_359	354_359	AX-116655949	2	83558137	4.99	0.892
BF_W	Global-GWAS	360_361	360_361	H3GA0006975	2	85151135	4.5	0.127
BF_W	LRFI-GWAS	360_361	360_361	AX-116175762	2	85277940	4.99	-0.176
LMCcalc	Global-GWAS	362_363	362_363	AX-116176302	2	87289847	4.51	0.885
BF_W	LRFI-GWAS	364_366	364_366	AX-116176833	2	89533075	4.56	-0.164
L*_GS	LRFI-GWAS	401_403	401_403	ASGA0011729	2	126745495	5.01	-1.557
pH24h_GS	LRFI-GWAS	401_403	401_403	AX-116186661	2	126981236	4.62	0.049
pH24h_GS	Global-GWAS	416_417	416_417	ALGA0118922	2	141103138	4.92	-0.052
a*_GS	LRFI-GWAS	422_424	422_424	AX-116716810	2	147951593	5.02	-0.712
DP	HRFI-GWAS	498_499	498_499	AX-116207811	3	71352397	4.55	0.006
DFI	LRFI-GWAS	553_555	553_555	AX-116219557	3	126619612	5.81	0.068
L*_GM	HRFI-GWAS	563_564	563_564	AX-116221734	4	3500258	5.16	2.327
pH24h_AD	Global-GWAS	568_569	568_570	AX-116222888	4	8866383	4.73	0.107
WHC	HRFI-GWAS	568_570	568_570	ALGA0022893	4	8926615	5.25	-18.732
L*_GM	HRFI-GWAS	583_584	583_584	AX-116727028	4	23188838	4.97	-1.075
L*_GM	HRFI-GWAS	590_591	590_591	AX-116227884	4	30422582	4.58	-1.031
L*_GM	HRFI-GWAS	592_594	592_594	AX-116228622	4	33168815	4.67	-1
ADG	HRFI-GWAS	614_615	614_615	AX-116233656	4	54840194	4.71	-0.033
b*_GM	Global-GWAS	625_626	625_626	AX-116236508	4	65719654	4.68	0.523
pH24h_GS	HRFI-GWAS	637_638	637_638	AX-116729133	4	77834969	4.59	-0.069
L*_GM	LRFI-GWAS	677_678	677_678	AX-116248908	4	118156608	4.59	-1.856
b*_GM	Global-GWAS	679_681	679_681	AX-116249145	4	119108681	4.51	-0.444
Shoulder_W	Global-GWAS	679_681	679_681	AX-116249351	4	120050234	4.53	-0.183
carcBFT	LRFI-GWAS	757_758	757_758	AX-116265244	5	66100317	7	-1.632
L*_GM	Global-GWAS	765_767	765_767	AX-116649188	5	74341682	6.36	-1.051
DP	Global-GWAS	772_773	771_774	WU_10.2_5_85211794	5	81443553	4.7	-0.005
b*_GS	HRFI-GWAS	771_774	771_774	AX-116751231	5	81910321	5.05	-1.317
DP	Global-GWAS	798_799	798_799	AX-116274437	6	2371001	4.8	0.005
b*_GS	Global-GWAS	798_799	798_799	AX-116274481	6	2568171	4.78	0.443
pH24h_LM	LRFI-GWAS	803_806	803_806	WU_10.2_6_7664397	6	7644906	4.98	-0.039
b*_GS	Global-GWAS	803_806	803_806	AX-116275510	6	8429547	5.72	0.605
b*_GS	HRFI-GWAS	803_806	803_806	AX-116275644	6	9133944	4.66	-0.614
L*_GS	HRFI-GWAS	803_806	803_806	AX-116275645	6	9136948	5.51	-1.625
L*_GS	Global-GWAS	803_806	803_806	AX-116782722	6	9150623	5.63	1.198
MQI	Global-GWAS	803_806	803_806	AX-116782722	6	9150623	7.5	-1.051
MQI	HRFI-GWAS	803_806	803_806	AX-116782722	6	9150623	4.9	-1.131
pH24h_SM	Global-GWAS	803_806	803_806	AX-116782722	6	9150623	6.12	-0.066
pH24h_GS	Global-GWAS	803_806	803_806	AX-116275650	6	9156412	5.67	-0.054
MQI	HRFI-GWAS	809_814	809_814	WU_10.2_6_13750573	6	13837643	4.6	-1.535
pH24h_SM	HRFI-GWAS	809_814	809_814	AX-116276661	6	14874409	4.87	0.106
b*_GS	HRFI-GWAS	809_814	809_814	AX-116660433	6	14915327	4.61	-0.484
WHC	HRFI-GWAS	809_814	809_814	H3GA0053145	6	16038522	5.17	16.64
pH24h_LM	HRFI-GWAS	809_814	809_814	AX-116636071	6	16079313	5.81	0.105
WHC	Global-GWAS	814_816	814_817	AX-116641374	6	18985936	5.52	-12.58
pH24h_LM	Global-GWAS	814_816	814_817	AX-116632418	6	19071992	7.7	-0.067
WHC	HRFI-GWAS	815_817	814_817	AX-116277294	6	19186264	4.66	-14.063
L*_GS	HRFI-GWAS	815_817	814_817	AX-116277701	6	20869239	4.81	-1.556
WHC	Global-GWAS	822_823	822_823	AX-116278977	6	26545011	4.77	-10.673
ADG	HRFI-GWAS	836_837	836_837	AX-116281217	6	40712495	5.54	0.029
a*_GS	HRFI-GWAS	877_878	877_878	AX-116653371	6	81434800	4.61	-0.425
FCR	HRFI-GWAS	982_983	982_983	AX-116790757	7	15310026	4.5	0.124
carcBFT	HRFI-GWAS	1010_1011	1010_1011	AX-116318689	7	43871954	5.27	1.367
BF_W	HRFI-GWAS	1012_1013	1012_1013	AX-116319076	7	45684415	4.59	-0.191
b*_GS	Global-GWAS	1015_1018	1015_1018	AX-116319907	7	49996822	5.42	-0.482
DFI	Global-GWAS	1035_1036	1035_1036	AX-116323269	7	68236808	4.63	0.067
b*_GM	LRFI-GWAS	1043_1044	1043_1044	AX-116849635	7	76536643	4.8	0.585
pH24h_AD	LRFI-GWAS	1053_1054	1053_1054	AX-116327117	7	86537327	4.57	0.18
carcBFT	Global-GWAS	1061_1062	1061_1062	AX-116329125	7	94311274	4.82	-1.071
FCR	Global-GWAS	1073_1075	1073_1076	AX-116767487	7	106200915	4.57	0.062
pH24h_AD	LRFI-GWAS	1074_1076	1073_1076	INRA0028058_60k	7	107229684	4.89	0.067
b*_GS	Global-GWAS	1073_1075	1073_1076	AX-116332019	7	107563992	4.52	0.364
L*_GS	HRFI-GWAS	1074_1076	1073_1076	ASGA0035957	7	108853820	4.54	3.795

pH24h_LM	HRFI-GWAS	1078_1079	1078_1079	AX-116763826	7	111787646	4.64	0.082
pH24h_LM	Global-GWAS	1080_1081	1080_1081	AX-116333311	7	113599753	4.95	-0.057
RFI	LRFI-GWAS	1084_1086	1084_1086	AX-116334309	7	117994063	4.54	-0.053
FCR	LRFI-GWAS	1084_1086	1084_1086	DBNP0002208	7	117999329	5.17	-0.14
carcBFT	LRFI-GWAS	1084_1086	1084_1086	AX-116849856	7	118636613	4.83	-2.347
L*_GM	HRFI-GWAS	1095_1096	1095_1096	AX-116337290	8	6636109	4.74	1.009
carcBFT	LRFI-GWAS	1108_1109	1108_1111	AX-116340278	8	19784122	4.67	1.06
pH24h_LM	Global-GWAS	1108_1110	1108_1111	AX-116340314	8	19936635	5.33	-0.052
L*_GS	Global-GWAS	1108_1110	1108_1111	AX-116340573	8	20939365	4.54	-2.093
pH24h_LM	LRFI-GWAS	1110_1111	1108_1111	AX-116340672	8	21349090	5.12	0.046
pH24h_LM	Global-GWAS	1112_1113	1112_1113	WU_10.2_8_23981291	8	23155364	5.22	0.052
pH24h_LM	LRFI-GWAS	1112_1113	1112_1113	AX-116341224	8	23466871	5.22	0.051
pH24h_LM	LRFI-GWAS	1192_1193	1192_1193	AX-116736551	8	103478967	4.55	-0.044
RFI	LRFI-GWAS	1199_1201	1199_1201	WU_10.2_8_119036157_80k	8	110934427	4.54	0.036
b*_GM	HRFI-GWAS	1210_1211	1210_1211	AX-116363406	8	121850337	4.52	-0.577
LMCcalc	LRFI-GWAS	1221_1223	1221_1223	AX-116365858	8	132849882	4.53	-1.126
FCR	LRFI-GWAS	1221_1223	1221_1223	AX-116366002	8	133403522	4.6	0.068
pH24h_LM	HRFI-GWAS	1229_1230	1229_1230	AX-116367719	9	1024688	4.74	0.065
RFI	Global-GWAS	1238_1239	1238_1239	AX-116369608	9	10080418	5.01	-0.039
pH24h_LM	LRFI-GWAS	1247_1249	1247_1249	DIAS0000480	9	20275683	4.83	-0.041
DFI	LRFI-GWAS	1295_1296	1295_1296	AX-116382987	9	67344507	4.71	-0.067
MQI	LRFI-GWAS	1347_1348	1347_1348	AX-116394499	9	119393699	4.95	2.952
pH24h_SM	LRFI-GWAS	1347_1348	1347_1348	AX-116394499	9	119393699	4.53	0.184
L*_GS	LRFI-GWAS	1349_1350	1349_1350	AX-116394883	9	121583592	4.54	1.104
Belly_W	LRFI-GWAS	1356_1357	1356_1357	WU_10.2_9_140618132_80k	9	128157351	4.63	-0.112
Loin_W	Global-GWAS	1395_1396	1395_1396	AX-116405696	10	28328451	4.96	-0.192
a*_GS	HRFI-GWAS	1397_1398	1397_1398	AX-116636342	10	29901587	4.7	0.447
a*_GS	HRFI-GWAS	1408_1409	1408_1409	AX-116408358	10	40398442	4.71	-0.497
Shoulder_W	LRFI-GWAS	1417_1418	1417_1418	AX-116410601	10	49720163	4.62	-0.56
WHC	HRFI-GWAS	1431_1437	1431_1437	AX-116414056	10	65058554	5.99	-22.915
a*_GS	LRFI-GWAS	1444_1446	1444_1446	WU_10.2_11_6479535_80k	11	6724010	4.81	-0.531
a*_GS	LRFI-GWAS	1449_1450	1449_1450	AX-116417976	11	11924688	5.55	-1.264
LMCcalc	LRFI-GWAS	1458_1460	1458_1460	AX-116419566	11	20359296	5	0.921
carcBFT	LRFI-GWAS	1458_1460	1458_1460	AX-116419728	11	21119127	5.3	-1.161
DP	LRFI-GWAS	1458_1460	1458_1460	H3GA0031530	11	21153806	5.26	-0.006
DP	LRFI-GWAS	1462_1463	1462_1463	AX-116420567	11	24874673	4.76	-0.006
DP	Global-GWAS	1464_1465	1464_1465	AX-116420903	11	26232809	4.87	-0.005
DP	LRFI-GWAS	1464_1465	1464_1465	AX-116420898	11	26251342	4.84	-0.005
DP	LRFI-GWAS	1483_1484	1483_1484	WU_10.2_11_50039446_80k	11	45544633	4.51	-0.006
DP	LRFI-GWAS	1485_1488	1485_1488	AX-116809574	11	48509307	5.09	0.005
pH24h_GS	LRFI-GWAS	1502_1503	1502_1503	AX-116430137	11	64108986	4.51	0.106
DFI	Global-GWAS	1502_1503	1502_1503	AX-116744122	11	64218801	5.45	0.073
ADG	HRFI-GWAS	1525_1527	1525_1527	AX-116435161	12	7621549	5.99	0.036
RFI	LRFI-GWAS	1525_1527	1525_1527	ALGA0114079	12	8907832	4.56	0.042
LMCcalc	Global-GWAS	1530_1533	1530_1533	AX-116436344	12	13180722	4.96	-1.031
a*_GS	Global-GWAS	1530_1533	1530_1533	AX-116793258	12	14706890	4.53	0.417
Loin_W	Global-GWAS	1540_1541	1540_1541	WU_10.2_12_22557810_80k	12	22192008	4.57	0.175
b*_GS	LRFI-GWAS	1581_1582	1581_1582	AX-116657858	13	1050752	4.87	-0.757
pH24h_AD	Global-GWAS	1583_1584	1583_1584	AX-116447665	13	3776387	4.64	0.116
a*_GM	Global-GWAS	1589_1590	1589_1590	AX-116449095	13	9908086	4.5	0.656
b*_GM	Global-GWAS	1596_1598	1596_1598	AX-116450762	13	16605006	4.5	-0.466
b*_GS	Global-GWAS	1596_1598	1596_1598	AX-116450988	13	17419221	4.72	0.429
b*_GS	Global-GWAS	1601_1603	1601_1603	WU_10.2_13_25061425	13	22889617	5.23	-0.456
Loin_W	LRFI-GWAS	1628_1631	1628_1631	AX-116458776	13	49176584	5.08	-0.186
pH24h_GS	Global-GWAS	1718_1719	1718_1719	DIAS0001553	13	138558004	5.32	0.126
pH24h_GS	Global-GWAS	1720_1722	1720_1722	AX-116852345	13	141001086	5.27	0.117
pH24h_SM	Global-GWAS	1720_1722	1720_1722	AX-116852345	13	141001086	5.19	0.132
RFI	LRFI-GWAS	1764_1765	1764_1765	AX-116489818	13	184597311	5.62	-0.033
b*_GS	HRFI-GWAS	1769_1770	1769_1770	ASGA0100757	13	189662482	4.55	0.461
L*_GS	HRFI-GWAS	1791_1792	1791_1792	AX-116759288	14	2562590	5.03	1.99
b*_GS	HRFI-GWAS	1791_1792	1791_1792	AX-116759288	14	2562590	4.68	0.695
a*_GM	LRFI-GWAS	1798_1799	1798_1799	AX-116496376	14	9268092	4.61	1.306
Ham_W	LRFI-GWAS	1800_1801	1800_1801	AX-116496934	14	11696294	4.68	0.148
DP	LRFI-GWAS	1802_1803	1802_1803	AX-116497437	14	13833569	5.04	0.005
DP	LRFI-GWAS	1804_1805	1804_1805	AX-116804942	14	15000871	5.67	0.005
Belly_W	LRFI-GWAS	1870_1872	1870_1872	AX-116514757	14	82409348	4.9	-0.175
L*_GS	LRFI-GWAS	1878_1879	1878_1879	AX-116516866	14	89992977	4.78	1.678
RFI	LRFI-GWAS	1900_1901	1900_1901	AX-116522437	14	111143930	4.96	-0.058
pH24h_AD	LRFI-GWAS	1902_1903	1902_1903	AX-116523140	14	113920119	4.52	0.071
RFI	Global-GWAS	1918_1920	1918_1920	AX-116527088	14	130819658	5.42	-0.052
MQI	Global-GWAS	1976_1977	1976_1977	AX-116756045	15	45371999	4.98	0.899
pH24h_SM	Global-GWAS	1976_1977	1976_1977	AX-116756045	15	45371999	5.88	0.068
pH24h_LM	LRFI-GWAS	1980_1981	1980_1981	AX-116540453	15	49329427	4.54	0.092
pH24h_SM	Global-GWAS	2011_2012	2011_2012	WU_10.2_15_89386471_80k	15	80076089	4.51	0.061
a*_GS	HRFI-GWAS	2048_2049	2048_2049	AX-116554203	15	117824259	4.68	-0.517
MQI	LRFI-GWAS	2083_2084	2083_2084	AX-116562203	16	11221032	4.54	0.743
RFI	LRFI-GWAS	2085_2086	2085_2086	AX-116562700	16	13519093	4.57	0.04
Loin_W	HRFI-GWAS	2102_2103	2102_2103	AX-116566955	16	30715316	4.55	0.284
pH24h_SM	LRFI-GWAS	2106_2107	2106_2107	AX-116567928	16	34093112	5.27	0.175
MQI	Global-GWAS	2148_2149	2148_2149	AX-116576954	16	76541834	5.09	-1.201
pH24h_GS	Global-GWAS	2148_2149	2148_2149	AX-116787882	16	76705149	4.54	-0.067
FCR	Global-GWAS	2159_2160	2159_2160	AX-116579415	17	7644014	4.76	0.071
RFI	Global-GWAS	2200_2201	2200_2201	WU_10.2_17_53675457	17	48029388	5.48	-0.04
b*_GS	Global-GWAS	2209_2210	2209_2210	AX-116591053	17	57309334	5.6	0.469
Shoulder_W	HRFI-GWAS	2215_2217	2215_2217	AX-116592317	18	274712	5.23	-0.255
a*_GM	Global-GWAS	2219_2220	2219_2220	AX-116647036	18	4732300	4.56	0.785
WHC	Global-GWAS	2219_2220	2219_2220	AX-116796099	18	4745191	5.24	15.046
WHC	HRFI-GWAS	2223_2225	2223_2225	AX-116594138	18	8857019	4.99	15.666
L*_GM	LRFI-GWAS	2256_2257	2256_2257	AX-116630408	18	41981051	4.52	-1.03

Annexe 2

TRAIT	SSC	POS_start	POS_end	Papier	Annee
CMJR	1	23	24	Bai	2017
CMJR	2	146	147	Bai	2017
CMJR	10	64	65	Bai	2017
CMJR	10	66	67	Bai	2017
CMJR	12	52	53	Bai	2017
CMJR	13	46	47	Bai	2017
CMJR	13	194	195	Bai	2017
CMJR	15	1	2	Bai	2017
CMJ	4	11	12	Ding	2017
CMJ	6	143	144	Ding	2017
CMJ	6	165	166	Ding	2017
CMJ	7	115	116	Ding	2017
IC	12	17	19	Ding	2017
CMJ	1	54	57	Do_1	2014
CMJR	1	26	28	Do_1	2014
CMJR	1	54	57	Do_1	2014
CMJR	9	109	111	Do_1	2014
CMJR	13	199	202	Do_1	2014
CMJR	13	203	205	Do_1	2014
GMQ	17	55	57	Do_1	2014
CMJR	3	89	90	Do_2	2014
CMJR	7	17	18	Do_2	2014
CMJR	7	91	92	Do_2	2014
CMJR	8	81	83	Do_2	2014
CMJR	15	73	75	Do_2	2014
CMJR	17	40	41	Do_2	2014
CMJR	1	7	8	Do_2	2014
CMJR	7	17	18	Do_2	2014
CMJR	8	26	27	Do_2	2014
CMJR	8	82	83	Do_2	2014
CMJR	10	40	41	Do_2	2014
CMJR	14	129	130	Do_2	2014
CMJR	15	122	123	Do_2	2014
IC	1	75	76	Horodyska	2017
IC	4	79	82	Horodyska	2017
IC	6	84	85	Horodyska	2017
IC	6	86	87	Horodyska	2017
IC	15	134	135	Horodyska	2017
CMJ	1	149	150	Jiao	2014
CMJ	2	41	43	Jiao	2014
CMJ	8	73	75	Jiao	2014
CMJ	10	38	48	Jiao	2014
CMJ	10	66	68	Jiao	2014
CMJR	2	36	43	Jiao	2014
CMJR	4	5	7	Jiao	2014
GMQ	1	151	153	Jiao	2014
GMQ	4	5	7	Jiao	2014
GMQ	11	24	27	Jiao	2014
GMQ	14	17	19	Jiao	2014
IC	4	4	6	Jiao	2014
IC	7	119	121	Jiao	2014
IC	10	38	49	Jiao	2014
IC	11	5	7	Jiao	2014
IC	15	102	104	Jiao	2014

CMJ	1	149	151	Onteru	2013
CMJ	1	153	156	Onteru	2013
CMJ	1	158	160	Onteru	2013
CMJ	1	167	169	Onteru	2013
CMJ	4	76	77	Onteru	2013
CMJ	4	126	127	Onteru	2013
CMJ	6	146	147	Onteru	2013
CMJ	7	116	117	Onteru	2013
CMJ	8	28	29	Onteru	2013
CMJ	11	27	28	Onteru	2013
CMJ	14	13	14	Onteru	2013
CMJ	14	27	28	Onteru	2013
CMJ	14	52	53	Onteru	2013
CMJ	14	55	56	Onteru	2013
CMJ	14	57	58	Onteru	2013
CMJ	14	98	99	Onteru	2013
CMJ	14	116	117	Onteru	2013
CMJ	15	52	53	Onteru	2013
CMJ	17	56	57	Onteru	2013
CMJR	2	106	108	Onteru	2013
CMJR	3	18	19	Onteru	2013
CMJR	5	19	20	Onteru	2013
CMJR	7	15	16	Onteru	2013
CMJR	7	33	34	Onteru	2013
CMJR	8	15	16	Onteru	2013
CMJR	8	16	17	Onteru	2013
CMJR	8	135	136	Onteru	2013
CMJR	9	11	12	Onteru	2013
CMJR	9	122	123	Onteru	2013
CMJR	14	1	2	Onteru	2013
CMJR	14	3	4	Onteru	2013
CMJR	14	55	56	Onteru	2013
CMJR	14	82	84	Onteru	2013
CMJR	15	45	46	Onteru	2013
CMJR	17	55	56	Onteru	2013
GMQ	1	75	78	Onteru	2013
GMQ	1	101	103	Onteru	2013
GMQ	1	149	151	Onteru	2013
GMQ	1	153	156	Onteru	2013
GMQ	1	156	158	Onteru	2013
GMQ	1	158	160	Onteru	2013
GMQ	1	243	245	Onteru	2013
GMQ	2	136	138	Onteru	2013
GMQ	4	30	35	Onteru	2013

GMQ	4	80	81	Onteru	2013
GMQ	5	90	91	Onteru	2013
GMQ	5	93	94	Onteru	2013
GMQ	6	22	23	Onteru	2013
GMQ	6	54	55	Onteru	2013
GMQ	6	146	149	Onteru	2013
GMQ	7	87	88	Onteru	2013
GMQ	7	117	118	Onteru	2013
GMQ	9	21	22	Onteru	2013
GMQ	9	24	25	Onteru	2013
GMQ	10	12	13	Onteru	2013
GMQ	10	44	45	Onteru	2013
GMQ	11	27	28	Onteru	2013
GMQ	12	39	40	Onteru	2013
GMQ	13	19	20	Onteru	2013
GMQ	13	21	22	Onteru	2013
GMQ	13	32	33	Onteru	2013
GMQ	14	26	28	Onteru	2013
GMQ	14	29	30	Onteru	2013
GMQ	14	40	41	Onteru	2013
GMQ	14	89	90	Onteru	2013
GMQ	14	98	99	Onteru	2013
GMQ	14	116	117	Onteru	2013
GMQ	15	121	122	Onteru	2013
GMQ	16	54	55	Onteru	2013
GMQ	17	23	24	Onteru	2013
GMQ	1	272	273	Quan	2018
GMQ	1	274	275	Quan	2018
GMQ	5	30	31	Quan	2018
GMQ	6	132	133	Quan	2018
GMQ	9	117	118	Quan	2018
GMQ	13	206	207	Quan	2018
GMQ	16	84	85	Quan	2018
IC	2	139	141	Sahana	2013
IC	4	55	59	Sahana	2013
IC	4	91	93	Sahana	2013
IC	4	102	104	Sahana	2013
IC	5	12	15	Sahana	2013
IC	5	85	87	Sahana	2013
IC	7	1	3	Sahana	2013
IC	7	23	26	Sahana	2013
IC	7	45	47	Sahana	2013
IC	8	77	79	Sahana	2013
IC	9	54	56	Sahana	2013
IC	10	46	48	Sahana	2013
IC	14	3	5	Sahana	2013
IC	14	111	113	Sahana	2013
IC	15	12	14	Sahana	2013
IC	15	33	35	Sahana	2013
IC	15	40	43	Sahana	2013
IC	16	30	36	Sahana	2013
IC	16	72	78	Sahana	2013
IC	17	22	24	Sahana	2013
IC	18	51	53	Sahana	2013
CMJ	1	155	168	Silva	2019
GMQ	1	159	168	Silva	2019
GMQ	12	2	3	Silva	2019

Annexe 3

Aliakbari *et al. Genet Sel Evol* (2020) 52:57
https://doi.org/10.1186/s12711-020-00576-0

GSE Genetics Selection Evolution

RESEARCH ARTICLE Open Access

The impact of training on data from genetically-related lines on the accuracy of genomic predictions for feed efficiency traits in pigs

Amir Aliakbari*, Emilie Delpuech, Yann Labrune, Juliette Riquet and Hélène Gilbert

Abstract

Background: Most genomic predictions use a unique population that is split into a training and a validation set. However, genomic prediction using genetically heterogeneous training sets could provide more flexibility when constructing the training sets in small populations. The aim of our study was to investigate the potential of genomic prediction of feed efficiency related traits using training sets that combine animals from two different, but genetically-related lines. We compared realized prediction accuracy and prediction bias for different training set compositions for five production traits.

Results: Genomic breeding values (GEBV) were predicted using the single-step genomic best linear unbiased prediction method in six scenarios applied iteratively to two genetically-related lines (i.e. 12 scenarios). The objective for all scenarios was to predict GEBV of pigs in the last three generations (~400 pigs, G7 to G9) of a given line. For each line, a control scenario was set up with a training set that included only animals from that line (target line). For all traits, adding more animals from the other line to the training set did not increase prediction accuracy compared to the control scenario. A small decrease in prediction accuracies was found for average daily gain, backfat thickness, and daily feed intake as the number of animals from the target line decreased in the training set. Including more animals from the other line did not decrease prediction accuracy for feed conversion ratio and residual feed intake, which were both highly affected by selection within lines. However, prediction biases were systematic for these cases and might be reduced with bivariate analyses.

Conclusions: Our results show that genomic prediction using a training set that includes animals from genetically-related lines can be as accurate as genomic prediction using a training set from the target population. With combined reference sets, accuracy increased for traits that were highly affected by selection. Our results provide insights into the design of reference populations, especially to initiate genomic selection in small-sized lines, for which the number of historical samples is small and that are developed simultaneously. This applies especially to poultry and pig breeding and to other crossbreeding schemes.

Background

Given the large economic impact of feed efficiency in the swine industry, its evaluation requires accurate estimation of breeding values (BV) and selection of animals [1]. The most commonly used criterion to measure feed efficiency in livestock species is feed conversion ratio

*Correspondence: amir.alikbari@inrae.fr
GenPhySE, Université de Toulouse, INRAE, 31326 Castanet-Tolosan, France

(FCR) and is defined as feed intake per unit of live weight gain [2]. However, in 1963, residual feed intake (RFI) was introduced in cattle as an alternative criterion for feed efficiency [3]. In general, FCR and RFI are highly genetically correlated [4]. Nevertheless, selection of animals based on FCR can be accompanied by undesirable correlated responses in other traits such as appetite [5, 6], whereas selection for RFI is almost independent of these traits since RFI is feed intake adjusted for production trait by linear regression. Due to the high cost of measuring daily feed intake (DFI), and thus RFI and FCR [7], fewer phenotypic records are available, which reduces the accuracy of selection. Genomic selection has the potential to improve pig feed efficiency in some populations [8, 9]. Recent advances in genomic evaluation methodologies, such as single-step genomic best linear unbiased prediction (ssGBLUP), enable more accurate evaluations in small populations. The ssGBLUP combines phenotypic, genotypic, and pedigree information in a single genomic evaluation of animals [10–13]. The number of animals in the reference population has been shown to affect the accuracy of genomic predictions [14]. Multi-breed or admixed genomic evaluations have been proposed to increase the number of animals in reference sets for small populations [15], resulting in increases in prediction accuracy in some cases [16]. A study on multi-breed genomic evaluation using real data from Holstein and Jersey bulls showed that using a combined reference population resulted in comparable accuracies of genomic estimated breeding values (GEBV) in purebred validation sets, or exceeded that achieved with a purebred reference population of the same breed [17]. Adding a smaller population, i.e. Brown Swiss, to a reference population of Holstein and Jersey bulls resulted in slight increases in accuracy of predictions when breeds were considered as a single, joint population, while slight increases in accuracy were also observed if the breeds were treated as genetically-related traits [18]. Simulation studies with mixed reference populations also showed increases in prediction accuracy. A simulation study on genomic prediction across multiple populations in cattle showed that adding relatively few individuals from another population to a training set substantially increased the accuracy of predictions in the first population, regardless of the heritability (h^2) or marker density [19]. Another simulation study reported that genomic predictions using a combined versus a single reference population increased the accuracy of genomic predictions by 25%, with traits with a lower heritability benefiting more from the combination of populations [20]. However, using a combined reference population can be challenging if relationships between populations are absent: allele frequencies at the marker and/or causal loci, or causal variants themselves, can differ between populations, [15, 16]. Another limitation for across-breed genomic prediction is the inconsistency of linkage disequilibrium (LD) between markers and quantitative trait loci (QTL) between breeds, which is one of the assumptions of most genomic prediction models [17].

Given the presence of (ancestral) relationships between animals and the greater consistency of LD between genetically-related lines within a breed than between breeds that have been separated for decades, using a multi-line reference population may be more beneficial than using a multi-breed reference population [16]. However, the changes in allele frequency since separation of the lines may still represent a challenge for using a multi-line reference population [21]. To the best of our knowledge, the use of a multi-line genomic evaluation strategy in small, related lines using real data has not been studied, in spite of the existence of numerous related lines worldwide. Our hypothesis was that, in small porcine populations with few available ancestral samples, i.e. for which it is not possible to build large reference populations, including information from a genetically-related line in the training population could provide similar prediction accuracies as a within-line training population. Therefore, we explored reference populations with different structures that combined data from two lines that descended from a common origin, and compared the prediction accuracy obtained with that obtained when only information from the target line was used for training.

Methods

Population and data structure

The data were collected during a selection experiment that was conducted at INRAE (UE GenESI, Surgères, France, https://doi.org/10.15454/1.5572415481185847E12) on French Large White pigs. Two lines were established by nine generations of divergent selection for RFI from 2000 to 2015 [22]. The G0 generation resulted from the mating of 30 boars and 30 gilts from generation F0 using artificial insemination. Among the G0 animals, 116 boar candidates for selection from all 30 litters were tested for RFI to select six extreme founder boars for each line (LRFI: low RFI, and HRFI: high RFI). The two lines were initiated by mating the selected boars to about 35 random G0 gilts per line. Inbreeding was minimized at each generation. The development of each line continued with the selection of six boars out of 96 tested candidates in each generation from G1 to G9. In each generation, at least one additional parity was produced to evaluate correlated responses to selection for production traits on both females and castrated males (henceforth referred to as response animals). Selection candidates were evaluated

for RFI from 35 to 95 kg of body weight (BW), and response animals were evaluated from 10 weeks of age until slaughter (105 kg BW until G5 and 115 kg BW from G6 onwards). Animals were raised in four pens per batch and at least four batches per generation. Test pens were equipped with single-place electronic feeders ACEMA64 (ACEMO, France). Animals were offered ad libitum access to a pelleted diet based on cereals and soya bean meal containing 10 MJ net energy (NE)/kg and 160 g CP/kg, with a minimum of 0.80 g digestible Lys/MJ NE. In each generation, boars were selected based on a fixed RFI selection index that was established from pre-computed phenotypic correlations between DFI (g/day) and average daily gain (ADG, g/day) between 35 and 95 kg BW, and live backfat thickness (BFT, mm) at 95 kg BW [23], as RFI $= \text{DFI} - 1.06 \times \text{ADG} - 37 \times \text{BFT}$. The average metabolic BW (AMBW) was the same for all selection candidates and therefore excluded from the selection index equation. Selection candidates had records for feed intake, body weight, and live body composition traits. In addition to these phenotypes, gilts and castrated males had records for carcass composition traits [23]. For the present study, RFI, FCR, DFI, ADG and BFT were analyzed. These traits were available for both selection candidates and response animals. The number of observations for the five traits for each line are in Table 1. RFI of selection candidates was computed between 35 and 95 kg BW as the residual of a multiple linear regression of DFI on the traits included in the selection index. For gilts and castrated males from the correlated response batches, RFI was estimated from 10 weeks of age to slaughter as the residual of a multiple linear regression of DFI on AMBW, ADG from 10 weeks of age to slaughter, carcass BFT (carcBFT), and lean meat content (LMC; computed from cut weights) at slaughter. AMBW was included to account for maintenance requirements and the other traits were included to account for production requirements. [22]. Fixed effects included in the regression model to compute RFI of response animals were sex, pen size, contemporary group and BW at the beginning of the test. Complete pedigree information was collected from F0 to G9, plus up to 10 generations of ancestors, and contained 7046 animals (Table 1).

Combining and standardizing traits

Preliminary analyses on the five traits showed high genetic correlations between similar traits measured in selection candidate and response animals ($>0.80 \pm 0.11$, except 0.75 ± 0.08 between live BFT and carcass BFT). Therefore, to increase the amount of information, corresponding traits in selection candidate and response animals were combined for further analyses. Since animals differed in age and BW when measurements were taken, for each trait, records from selection candidates were standardized to the variance of the corresponding trait in the response animals as:

$$y_{Rij} = \frac{y_{sij}}{\sigma_{si}} \sigma_{Ri}$$

where y_{Rij} is the standardized trait i ($i = 1,\ldots 5$) for selection candidate j, y_{sij} is the record of trait i measured on animal j, σ_{si} is the phenotypic standard deviation of trait i measured on selection candidates, and σ_{Ri} is the phenotypic standard deviation of trait i measured on females and castrated males in the response batches. Descriptive statistics of these traits are in Table 2.

Single nucleotide polymorphism (SNP) genotyping data and imputation

SNP genotyping data were available for all selected boars and their mates from G0 to G9, additional pigs from response batches of G6 to G8, and all selection candidates in G9. In total, 1647 animals had SNP genotypes, of which 286 animals were genotyped with the Porcine SNP60v2 BeadChip (Illumina) (64,232 SNPs) and 1361 animals with the GGP Porcine HD Array (Illumina) (68,516 SNPs). Genotype quality control excluded SNPs with a call rate lower than 95%, individuals with a call rate lower than 90%, SNPs that were not in Hardy–Weinberg equilibrium ($p < 10^{-10}$), SNPs with a minor allele frequency lower than 0.01, and individuals with parent–offspring incompatibility (e.g., opposite homozygotes) with at least one parent. The PLINK software was used for SNP and individual genotype quality control [24]. SNPs on the sex chromosomes were removed. After quality control of each SNP chip dataset, the SNPs present in each panel were imputed to the alternative panel using the FImpute software [25] in a single step. The two SNP chips shared 42,800 SNPs. The number of genotyped animals retained after imputation was 1643, and the final genotype dataset contained 64,233 informative SNPs. Thus, all animals had equal genotypic information. Genotypes were coded as 0, 1, or 2 for later calculation of the genomic relationship matrix. The number of animals with genotype data per generation and line is in Table 1.

Model and analyses

Predictions obtained with BLUP are based on the assumption of no genetic differences between subpopulations [26, 27]. Therefore, to account for selection in our dataset, all genetic and genomic analyses were carried out with bivariate approaches, i.e. the five other traits were individually paired with the selection index in two-trait model analyses. By including the selection criterion, the analyses of other

Table 1 Numbers of animals in the pedigree and data structure

	Ancestors	F0	G0	HRFI										
				G0	G1	G2	G3	G4	G5	G6	G7	G8	G9	Total
Pedigree	159	67	104	48	216	297	277	260	270	795	474	292	280	3209
Pedigree only				1	2	89	78	62	68	352	149	5	0	806
Pedigree and genotype only				41	41	42	44	36	47	40	35	42	91	459
ADG														
Phenotype only				0	167	160	149	156	149	304	194	148	93	1520
Phenotype and genotype				6	6	6	6	6	6	71	73	66	92	338
Missing				0	0	0	0	0	0	28	23	31	4	86
BFT														
Phenotype only				0	167	160	149	156	149	237	176	62	84	1340
Phenotype and genotype				6	6	6	6	6	6	71	73	66	92	338
Missing				0	0	0	0	0	0	95	41	117	13	266
DFI														
Phenotype only				0	166	160	149	156	149	263	182	138	93	1456
Phenotype and genotype				6	6	6	6	6	6	71	73	66	92	338
Missing				0	1	0	0	0	0	69	35	41	4	150
FCR														
Phenotype only				0	166	160	148	156	149	263	182	138	93	1455
Phenotype and genotype				4	6	6	6	6	6	71	73	66	92	336
Missing				2	1	0	1	0	0	69	35	41	4	153
RFI														
Phenotype only				0	164	159	146	156	143	185	147	56	80	1236
Phenotype and genotype				6	6	6	6	6	6	71	73	66	92	338
Missing				0	3	1	3	0	6	147	70	123	17	370
	Ancestors	**0**	**G0**	**LRFI**										
				G0	**G1**	**G2**	**G3**	**G4**	**G5**	**G6**	**G7**	**G8**	**G9**	**Total**
Pedigree	159	67	104	46	203	303	314	327	357	826	481	344	280	3481
Pedigree only				0	1	98	100	107	130	337	132	8	0	913
Pedigree and genotype only				40	35	40	41	43	43	48	55	48	93	486
ADG														
Phenotype only				0	161	159	167	171	178	359	211	203	95	1704
Phenotype and genotype				6	6	6	6	6	6	74	73	74	90	347
Missing				0	0	0	0	0	0	8	10	11	2	31
BFT														
Phenotype only				0	161	159	167	171	178	284	206	105	86	1517
Phenotype and genotype				6	6	6	6	6	6	74	73	74	90	347
Missing				0	0	0	0	0	83	15	109	1	1	218
DFI														
Phenotype only				0	160	159	167	171	178	316	206	194	95	1646
Phenotype and genotype				6	6	6	6	6	6	74	73	74	90	347
Missing				0	1	0	0	0	0	51	15	20	2	89
FCR														
Phenotype only				0	159	159	167	171	178	316	208	195	95	1648
Phenotype and genotype				6	6	6	6	6	74	73	74	90		347
Missing				0	2	0	0	0	0	51	13	19	2	87
RFI														
Phenotype only				0	160	158	161	171	173	230	165	101	80	1399
Phenotype and genotype				6	6	6	6	6	6	74	73	74	90	347
Missing				0	1	1	6	0	5	137	56	113	17	336

HRFI high RFI line, *LRFI* low RFI line, *Ancestors* animals before the base generation, *F0* base generation, *G0 to G9* generations of selection 0 to 9, *RFI* residual feed intake, *ADG* average daily gain, *FCR* feed conversion ratio, *DFI* daily feed intake, *BFT* backfat thickness

Table 2 Descriptive statistics of the data for the studied traits in the HRFI and LRFI lines

Line	Trait	Number of records	Minimum	Maximum	Average	Coefficient of variation
HRFI	ADG	1868	0.44	1.07	0.76	11.03
	BFT	1687	9.67	49.27	27.33	26.62
	DFI	1802	1.37	3.20	2.18	12.54
	FCR	1799	2.13	3.81	2.8	9.26
	RFI	1581	−0.29	0.86	0.05	–
LRFI	ADG	2053	0.45	1.06	0.76	10.69
	BFT	1866	10.00	44.63	26.45	24.60
	DFI	1995	1.05	2.92	2.01	12.91
	FCR	1997	1.72	3.70	2.60	9.11
	RFI	1748	−0.56	0.46	−0.04	–

HRFI high RFI line, *LRFI* low RFI line, *ADG* average daily gain (kg/day), *BFT* backfat thickness (mm), *DFI* daily feed intake (kg/day), *FCR* feed conversion ratio (kg/kg), *RFI* residual feed intake (kg/day)

traits are conditioned based on all the information that was used for selection [28–30].

Preliminary analyses were carried out using a general linear model in R (glm procedure) to evaluate the significance ($p<0.05$) of fixed environmental sources of variation. The significant fixed factors included pen size (5 levels: 8, 9, 10, 11, 12 pigs per pen), herd of birth (2 levels), sex (3 levels), and contemporary groups (CG, 99 levels). BW at slaughter was fitted in the model as a covariate only for BFT. CG were defined as animals born in the same week and raised in the same enclosure. Litter was fitted as a random environmental source of variation and its significance at the 5% level was determined using a likelihood ratio test.

The genetic analyses were performed using the AIREMLF90 and BLUPF90 software [31] for the BLUP and ssGBLUP methods, respectively. Prior to ssGBLUP evaluations, the variance components of the traits were obtained using the restricted maximum likelihood algorithm implemented in AIREMLF90. These analyses were performed using all available data and only the full pedigree relationship matrix (**A**). Variance components were estimated with the bivariate animal mixed model as follows:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z}_1\mathbf{a} + \mathbf{Z}_2\mathbf{l} + \mathbf{e}$$

where **y** is the vector of observations for the index and one of the five studied traits, **b** is the vector of fixed effects (described above), **a** is the vector of additive genetic effects, **l** is the vector of litter effects, and **e** is the vector of random residuals. **X**, $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are the incidence matrices for **b**, **a**, and **l**, respectively. Distributions assumed for the random terms are $\mathbf{a} \sim N(\mathbf{0}, \mathbf{G}_0 \otimes \mathbf{A})$, $\mathbf{l} \sim N(\mathbf{0}, \mathbf{R}_\mathbf{l} \otimes \mathbf{I})$, $\mathbf{e} \sim N(\mathbf{0}, \mathbf{R}_\mathbf{e} \otimes \mathbf{I})$, where $\mathbf{G}_0$ is a 2×2 symmetric (co)variance matrix of direct additive genetic effects, and $\mathbf{R}_\mathbf{l}$ and $\mathbf{R}_\mathbf{e}$ are 2×2 symmetric (co)variances matrices of litter and residual effects, respectively. **I** denotes the identity matrix.

Genomic breeding values were estimated using ssGBLUP with the same models in the BLUPF90 software, with the previously estimated (co)variances and using the **H** matrix, which is a combined relationship matrix of the **A** matrix and marker-based relationship matrix (**G**) of genotyped animals [10, 12]. The **G** matrix was constructed and scaled by $2\sum\{p_i(1-p_i)\}$, where p_i is the frequency of the second allele at locus i, following VanRaden [32]. Computation of the **H** matrices used outputs of BLUPF90 (**G**) and the full **A** matrix, which was obtained using the AGHmatrix R package [33]. In all scenarios, **G** had similar average diagonal elements as the pedigree relationship matrix for the genotyped animals ($\mathbf{A}_{22}$).

Scenarios

Two symmetric series of six scenarios, one for each line, were defined for genomic prediction. An overview of the scenarios is shown in Fig. 1. In all scenarios, genotyped animals of the last three generations (G7 to G9, 433 pigs for the LRFI and 399 pigs for the HRFI line) were considered for validation in a given line (target line), and their information was removed from the training dataset.

The training sets were structured based on which generations and line were used. Scenario 1 comprised only animals from the target line and was the control scenario since it represented a routine genomic prediction design where all data would be available from the same line. All other scenarios were compared to this control scenario to evaluate which combination of training populations from the two lines achieved a prediction accuracy similar to the control scenario. Scenarios 2 and 3 included the training set of scenario 1 and in addition, either the

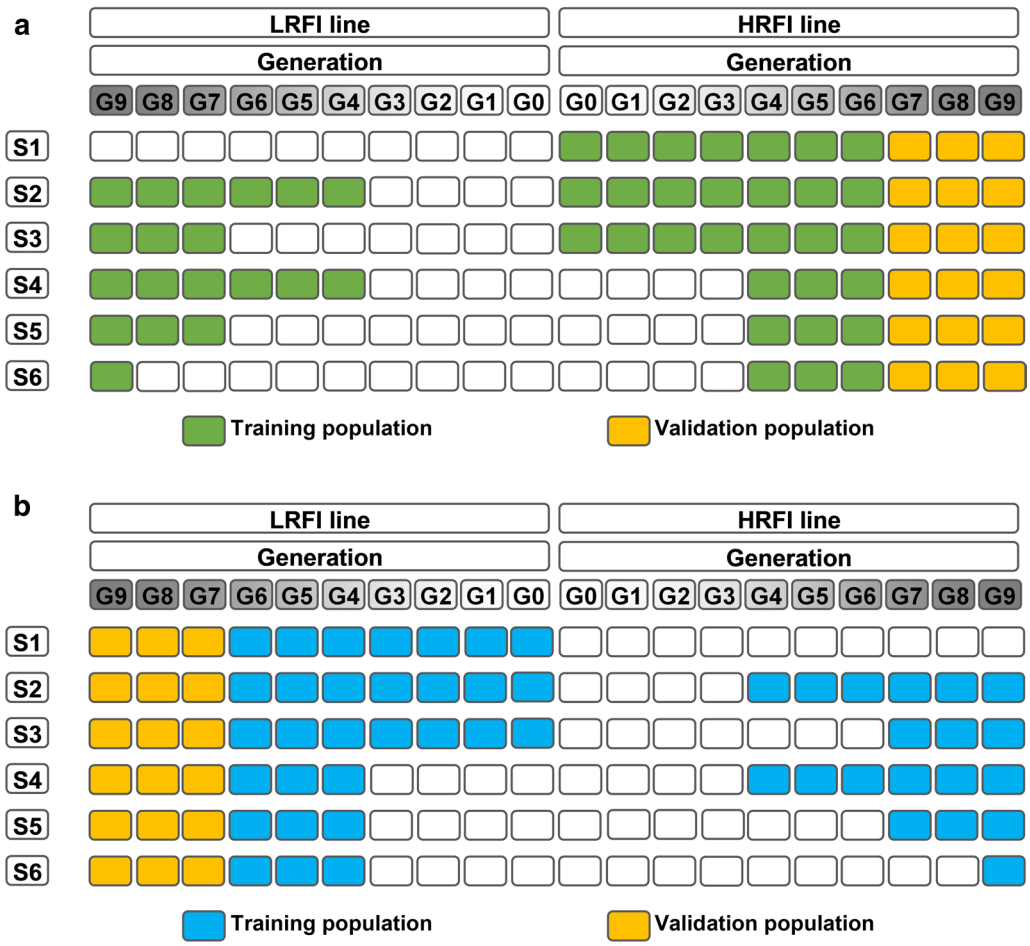

Fig. 1 Design of scenarios to predict validation animals in HRFI (**a**) and LRFI (**b**) lines

animals from G4 to G9 (scenario 2), or G7 to G9 (scenario 3) of the other line.

For scenarios 4 to 6, animals from the target line in the training set were limited to the three generations nearest to the validation set (G4 to G6). In scenarios 4 and 5, the contribution to the training set of the animals from the other line was as in scenario 2 (G4 to G9) and scenario 3 (G7 to G9), respectively. For scenario 6, the number of animals in the training set was close to that of scenario 1 and only animals from the G9 generation of the other line. Performance data of animals from the generation and line combinations that did not contribute to the training or validation sets were removed from the analysis, but their pedigree information was kept in order to trace relationships back to the founding generation. For example, phenotypes and genotypes of animals from G0 to G3 of both lines were removed for scenario 4, since they were not part of the training or validation sets. The number of genotyped animals in the training and validation sets for the 12 scenarios are in Table 3.

Accuracy and bias of genomic predictions

Usually the correlation between the vector of estimated breeding values (**EBV**) to be evaluated and the vector of true breeding values (**TBV**), $r(\mathbf{TBV}, \mathbf{EBV})$, cannot be computed. In the literature, multiple criteria have been proposed to quantify and compare prediction accuracies of genomic predictions between training and validation set structures and between prediction methods. Cross-validation approaches are often conducted based on $r(\mathbf{EBV}, \mathbf{y^*})$, where $\mathbf{y^*}$ is either the vector of phenotypes adjusted for fixed effects or the vector of deregressed EBV of the validation set. Thus, a widely used criterion is $r(\mathbf{EBV}, \mathbf{y^*})/\sqrt{h^2}$, where h^2 is the heritability of the trait. However, this criterion requires all the genotyped animals to have a sufficiently accurate $\mathbf{y^*}$ value [34]. When $\mathbf{y^*}$ is an adjusted phenotype of the animal's own measurement, it suffers from the inability to adjust for the random residual effects. In the optimum situation, the expected value of the correlation would then be the square root of heritability [35]. Alternatively, using an

Table 3 Number of genotyped animals in the training and validation sets for the six scenarios for the HRFI and LRFI validation sets

	HRFI		LRFI	
	Training	Validation	Training	Validation
Scenario 1	398	399	400	433
Scenario 2	1051	399	1005	433
Scenario 3	831	399	799	433
Scenario 4	859	399	825	433
Scenario 5	639	399	619	433
Scenario 6	389	399	403	433

HRFI high RFI line, *LRFI* low RFI line

Table 4 Estimates of variance components (SE) of the studied traits

Trait	Phenotypic variance	Heritability	Litter effects[a]
ADG	5811.70 (164.75)	0.25 (0.04)	0.10 (0.02)
BFT	14.37 (0.47)	0.36 (0.05)	0.12 (0.02)
DFI	0.04 (0.001)	0.24 (0.04)	0.09 (0.02)
FCR	0.04 (0.001)	0.24 (0.04)	0.07 (0.02)
RFI	0.01 (0.004)	0.12 (0.02)	0.08 (0.02)

ADG average daily gain (g/day), *BFT* backfat thickness (mm), *DFI* daily feed intake (kg/day), *FCR* feed conversion ratio (kg/kg), *RFI* residual feed intake (kg/day)

[a] As a proportion of phenotypic variance

EBV obtained from a complete dataset as the best predictor of TBV would cause autocorrelation between the reference and evaluated EBV when the training and validation sets are closely related through the pedigree, leading to higher correlations [35]. Legarra and Reverter [34] proposed to complement the cross-validation approach with a semi-parametric approach that can be used in a large number of cases, with the advantage of not requiring knowledge of the TBV or adjustment of phenotypes. The underlying assumptions of this approach are (1) the variance components are similar in the training and validation datasets, and (2) the validation set is sufficiently diverse and large (i.e. composed of various families). In brief, with their approach, the correlation between EBV using part of the dataset (partial) and EBV obtained using the whole dataset results in an estimator of the ratio of the accuracies of the EBV from these two datasets. We followed this approach to evaluate the potential for genomic prediction when including data from a related line compared to genomic prediction using all data from the target line, which will be referred to as $GEBV_w$ (GEBV obtained using the whole dataset), i.e., to obtain $GEBV_w$ for the validation set of each line, two separate ssGBLUP analyses were performed (one per line). $GEBV_p$ (GEBV obtained using partial dataset) were the GEBV obtained from the six scenarios for the validation sets in each target line. The criterion for prediction accuracy for each trait and each scenario was then the correlation between $GEBV_p$ and $GEBV_w$, $r(GEBV_p, GEBV_w)$. Bias of the genomic predictions was computed as the deviation of the regression coefficient of $GEBV_w$ on $GEBV_p$ from 1, as also proposed in [34].

Standard errors of the prediction accuracy correlations, r, were obtained as $\sqrt{[(1-r^2)/(n-2)]}$, where n is the number of animals used to obtain correlations in the validation sets. Differences between correlations in different scenarios were tested using the Williams t-test in the psych R package [36–38]. Significant differences between each scenario and the control scenario (scenario 1) are reported to identify the scenarios that provide prediction accuracies similar to the control scenario.

Relationships between training and validation sets

For each scenario, the maximum, average, and minimum relationship coefficients between training and validation sets in the **H** matrix were computed. To distinguish the strength of relationships originating from the two lines, all three measurements were computed separately for pigs of the validation set with the subset of the training set that belonged to (1) the target line and (2) the other line. The average relationships were calculated as the mean of the off-diagonal elements of the corresponding relationship matrices for the genotyped individuals.

Results

Variance components

The five studied traits showed low to moderate heritabilities that ranged from 0.12 ± 0.02 (RFI) to 0.36 ± 0.05 (BFT) (Table 4). The ratio of litter effect variance to phenotypic variance (l^2) was lower than the heritability for all traits, ranging from 0.07 ± 0.02 (FCR) to 0.12 ± 0.02 (BFT).

Prediction accuracies

Prediction accuracies, $r(GEBV_p, GEBV_w)$, for the different scenarios are shown in Fig. 2 for the two lines. Accuracies ranged from 0.07 to 0.73, depending on the validation line, trait, and scenario. The tested scenarios could be classified into two groups based on their design and how it affected the prediction accuracy of each trait. Removing the earlier generations of the target line from the training set (from scenarios 1, 2, 3 to scenarios 4, 5, 6) tended to decrease the prediction accuracy for ADG, BFT, and DFI, while FCR and RFI showed different patterns in response to changes in the structure of the training set.

The differences in prediction accuracies for ADG, BFT and DFI from scenario 1 to scenario 2 and 3 showed that the inclusion of different generations of the other line in the training set led to marginal changes in accuracy, with decreased correlations in most cases (BFT in the HRFI line and DFI). In scenarios 4, 5, and 6, the proportion of animals from the target line was low in the training set compared to scenarios 1, 2, and 3. This reduction generally led to a decrease in the prediction accuracies for ADG, BFT, and DFI compared to scenario 1. However, these differences in accuracy were only significant for ADG and BFT in the HRFI line and for DFI in the LRFI line.

Scenarios for FCR and RFI showed different patterns compared to the previous traits. Prediction accuracies for FCR followed a pattern similar to those of the other traits for all scenarios, except for scenario 3, which showed a 17 to 21% higher accuracy compared to scenario 1. Prediction accuracies for RFI decreased from scenario 1 to scenario 2, and scenario 1 to scenario 4 for the LRFI target line, which were the scenarios with the maximum number of individuals from the other line in the training set. In the other scenarios, the prediction accuracies for RFI were similar or higher than for scenario 1.

The prediction accuracies for FCR in all scenarios, except scenario 6, were higher for validation animals in the HRFI line than in the LRFI line. The average differences in accuracy by trait ranged from +0.07 for ADG to +0.40 for RFI (Fig.2).

Prediction biases

Overall, regression coefficients of $GEBV_w$ on $GEBV_p$ were consistently below 1 for FCR and RFI for both validation sets (Fig. 3). Regression coefficients for these two traits also showed more variation across the scenarios compared to ADG, BFT and DFI.

Bias for GEBV in the HRFI validation set followed the same trend, but at different magnitudes, for all traits,

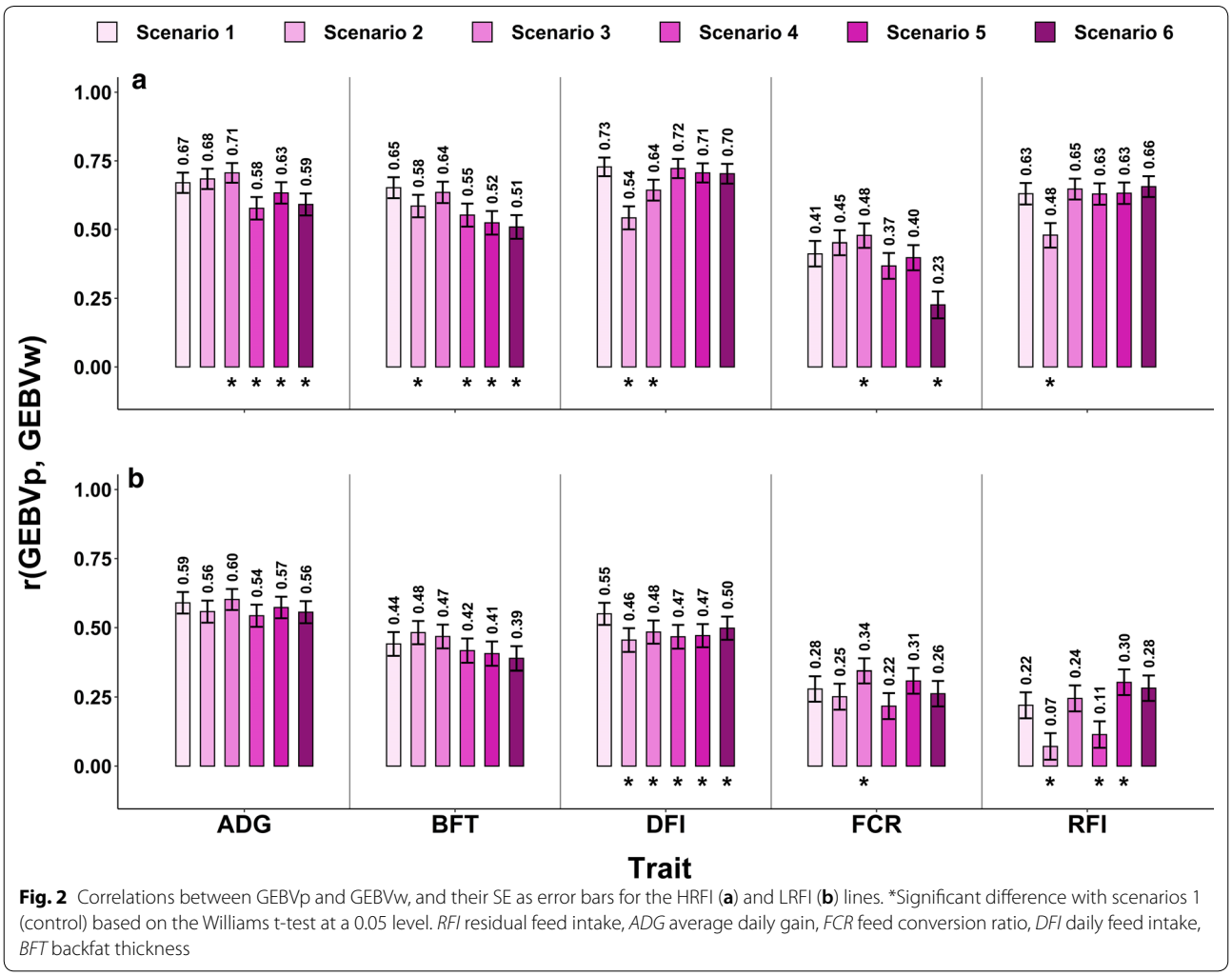

Fig. 2 Correlations between GEBVp and GEBVw, and their SE as error bars for the HRFI (**a**) and LRFI (**b**) lines. *Significant difference with scenarios 1 (control) based on the Williams t-test at a 0.05 level. *RFI* residual feed intake, *ADG* average daily gain, *FCR* feed conversion ratio, *DFI* daily feed intake, *BFT* backfat thickness

except ADG (Fig. 3a). On average, scenarios 1, 2, and 3 showed less biases than scenario 4, 5, and 6 for BFT, DFI, and FCR. The regression coefficient in scenario 1 was equal to 0.98 for RFI, slightly higher than 1 for BFT (1.08) and DFI (1.19), and lower than 1 for ADG (0.83) and FCR (0.74).

Prediction of GEBV for the LRFI validation set did not follow the same pattern of change across scenarios between the traits. Regression coefficients of all scenarios showed biases smaller than 1 for BFT, FCR, and RFI (Fig. 3b). Biases were smallest for DFI (scenario 6) and ADG (scenarios 1, 5 and 6). Overall, biases of GEBV for this line were moderate for scenario 6 compared to the other scenarios, except for BFT (0.53). Biases were larger for scenarios 2 and 4, compared to scenarios 5 and 6, for all traits except for BFT.

Relationships between and within training and validation sets

Relationships between the validation set and the training individuals from the target line were on average higher in scenarios 4 to 6 than in scenarios 1 to 3 (Fig. 4a and c). The highest average was obtained for scenario 4 (around 0.25) and the smallest average for scenarios 1 and 3 (around 0.16 and 0.17). The maximum relationship coefficient between these two cohorts was greater than 0.66 for all scenarios, with the smallest maximum found for scenario 1 when the training set included only individuals from the target line, and the highest maximum for scenario 4 (around 0.78), when the relative number of animals from the other line in the training set was larger.

Relationship coefficients between the validation set and the training individuals of the other line were lower than those with the training individuals of the target line, but the maximum values were reached for scenario 6, i.e. equal to 0.18 and 0.20 for the HRFI and LRFI target lines,

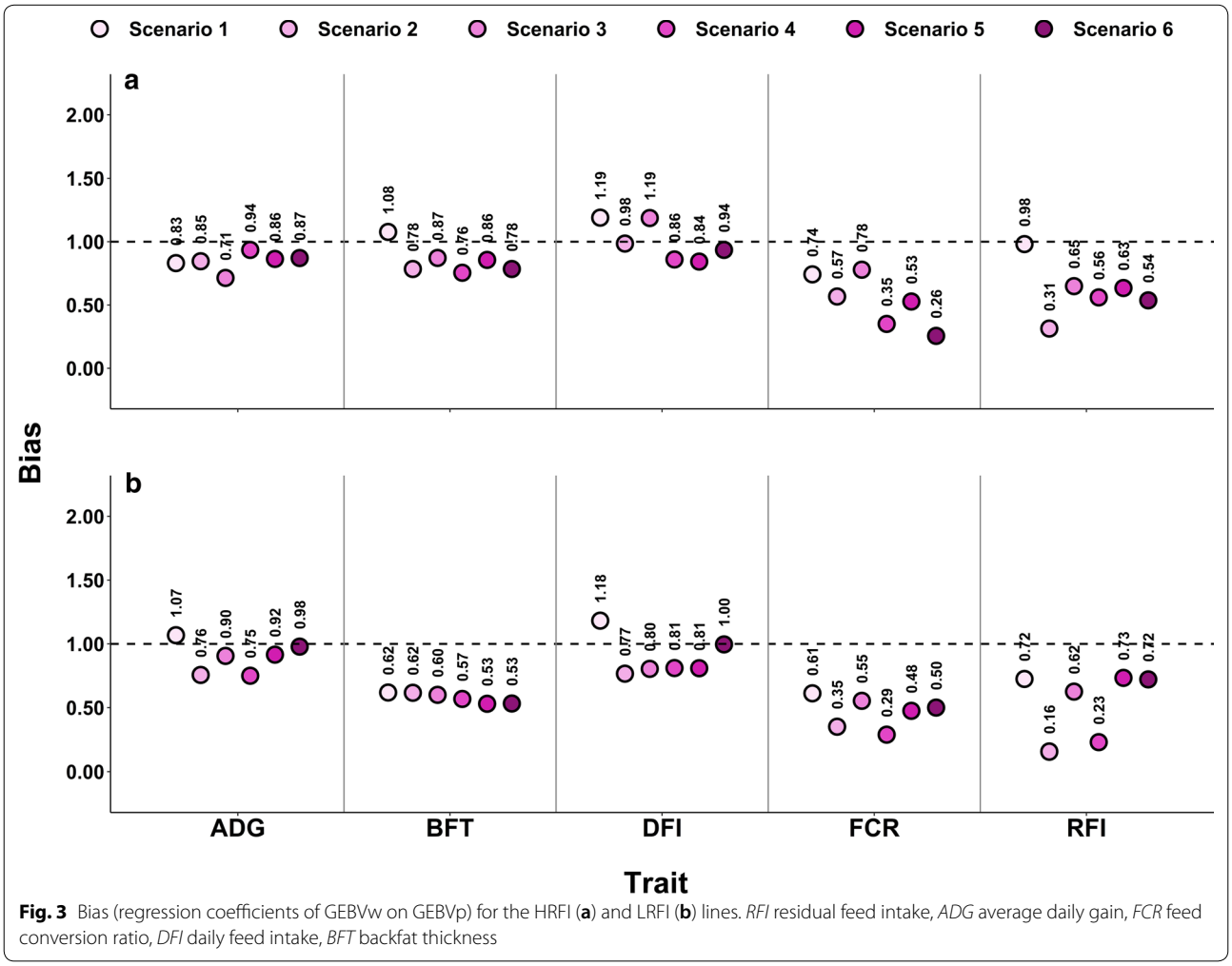

Fig. 3 Bias (regression coefficients of GEBVw on GEBVp) for the HRFI (**a**) and LRFI (**b**) lines. *RFI* residual feed intake, *ADG* average daily gain, *FCR* feed conversion ratio, *DFI* daily feed intake, *BFT* backfat thickness

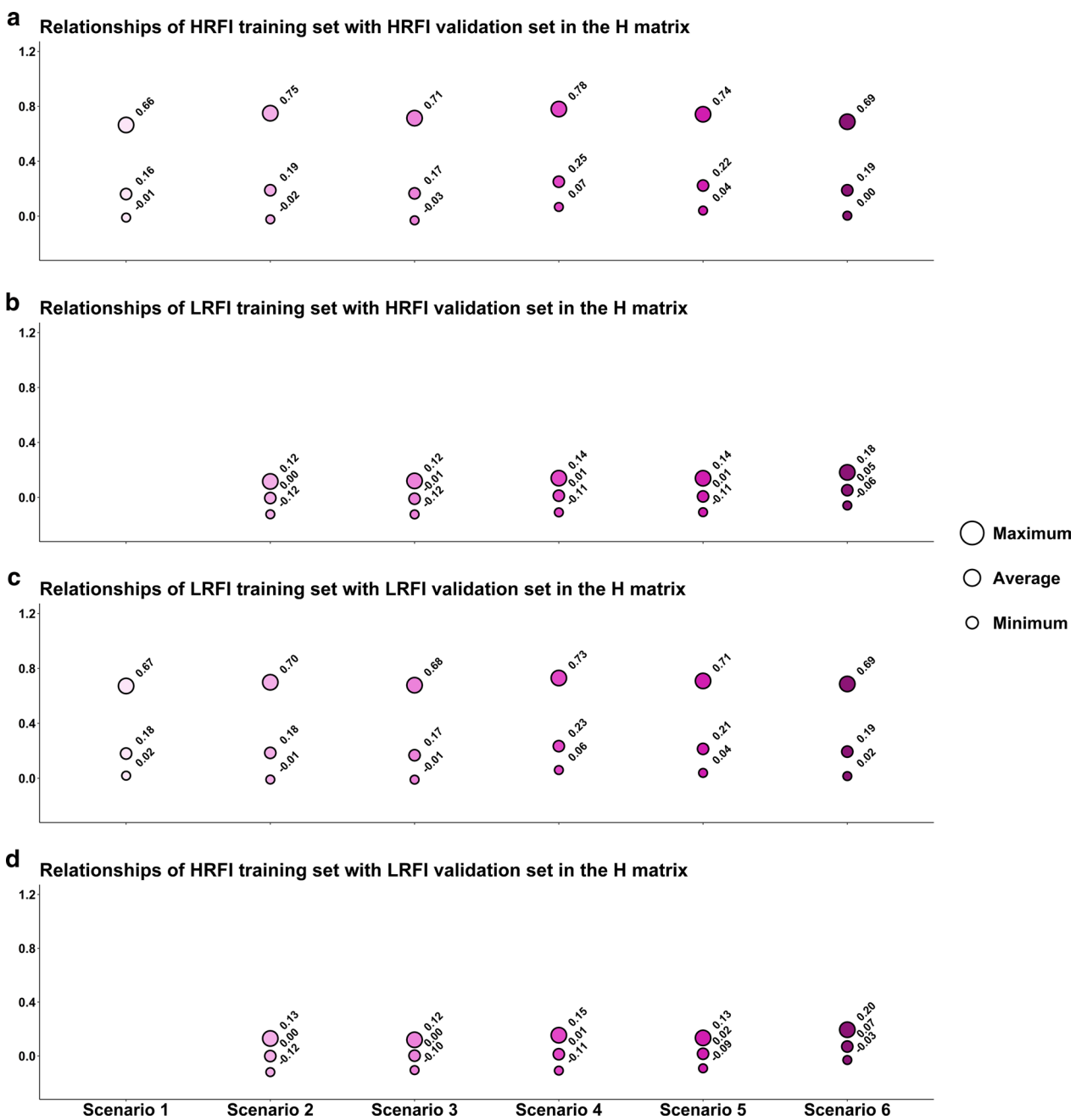

Fig. 4 Average, minimum and maximum relationship coefficients in the H matrix between individuals of the validation set, and individuals of the training set from the target line and from the reverse line, for **a** and **b** the HRFI target line, for **c** and **d** the LRFI target line

respectively (Fig. 4b and d). All other scenarios had lower maximum relationships, ranging from 0.12 to 0.15.

Discussion

The aim of our study was to investigate different combinations of two lines derived from a common origin to evaluate the potential of building a training set for the genomic prediction of feed efficiency related traits in lines that are small or do not have much data available. Multiplying by ~ 2.5 (scenario 2), ~ 2 (scenarios 3 and 4), and ~ 1.5 times (scenario 5) the number of genotyped individuals in the training set by recruiting animals from the other line show no or little increase of prediction accuracy. This would probably not justify the

additional genotyping costs involved. However, they can be considered for practical implementation of combined training sets since, in most cases, the prediction accuracies obtained in scenarios 5 and 6 were similar to those of the control scenario 1. These scenarios reflect most of the practical situations targeted in our study. Indeed, for breeding programs in small populations, phenotypic or genotypic information of individuals from earlier generations might not be available, and the sampling size in recent generations might be limited to a few hundred. Our results show that, a training population that includes recent generations of one population and data from a more distant subpopulation, could be a solution to achieve prediction accuracies similar to what would be achieved if data were available for individuals of the same population. This could even improve the prediction accuracies for traits under selection.

Computation of prediction accuracies and biases

Variance components of the evaluated traits were estimated using the **A** matrix on the full dataset with both lines combined. All estimated heritabilities were in the range of values reported in the literature for these traits [8, 39–42]. Using these variance components, the accuracy of GEBV was computed for the six scenarios to predict validation animals from each line using ssGBLUP. Prediction accuracies were computed using a cross-validation method combined with a semi-parametric approach [30]. Indeed, in our case, accuracies of the adjusted phenotypes or of deregressed EBV were too low to be used in a criterion such as $r(GEBV_p, y^*)/\sqrt{h^2}$, since only two-thirds of the individuals had their own phenotype. This would result in larger standard errors of the correlations and, thus, less power to test differences between scenarios, as shown in Additional file 1: Figures S1 and S2. The underlying assumptions of the semi-parametric approach are that (1) the validation set is sufficiently diverse and large (i.e. composed of various families), and (2) variance components are similar in the training and validation datasets. The first assumption was well covered in our study, since all breeding individuals, plus some progeny of each family, were phenotyped and genotyped. The second assumption was potentially less covered, which could explain some of the biases in prediction observed. Indeed, when estimating variance components separately in the two lines, different residual variances were estimated for some traits, resulting in lower heritability estimates for DFI (24%), FCR (43%), and RFI (22%) in the LRFI line than in the HRFI line. Legarra and Reverter [30] indicated that inflation of predictions in one or the other dataset due to changes in variances can cause biased GEBV. Thus, we also tested the use of estimates of variance components from the target line for the GEBV predictions, but this resulted in increases in biases by 0.016 to 0.121 in all situations but one (results not shown). In practice, scaling the observations by the residual or phenotypic standard deviations, or accounting for the heterogeneity of residual variance across lines, could be considered to account for such differences, as proposed by Reverter et al. [43] for heterogeneous variances across herds. An alternative could be to run bivariate analyses to consider correlated traits in the two lines, instead of a single trait across the two lines. Nevertheless, in our populations, estimates of the genetic variance of RFI as the trait under selection were consistent over the nine generations in each line. Therefore, differences in observed accuracy and bias between lines could not be explained by the heterogeneity of the genetic variance over the nine generations for the trait under selection.

Prediction accuracies for production traits

Although production traits and ssGBLUP have been discussed in the literature, few investigations have analyzed such traits in pigs with this method. Therefore, in the discussion that follows, we refer to published genomic prediction studies on these traits that often use other methods. Our objective in this part is to validate the prediction accuracies obtained with scenario 1, in which the structure of the training population is close to those of previous studies. When comparing studies, it is worth noting that ssGBLUP generally has a higher accuracy than the usual GBLUP or Bayesian approaches that use only data of genotyped animals. Thus in theory, the comparisons should favor ssGBLUP approaches. However, most previous studies were based on prediction to a single generation of candidates, which could favor higher prediction accuracies. In spite of these differences, overall, our estimates were within the range of accuracies reported in the literature, except for FCR and RFI, for which accuracies were higher in the HRFI validation set and lower in the LRFI line than those reported in the literature. In an investigation on 8113 Danish Duroc pigs with 60K imputed SNP genotyping information, an $r(GEBV_p, y^*)/\sqrt{h^2}$ of 0.41 was reported for ADG [41]. In a study with 620 commercial boars, an $r(GEBV_p, y^*)/\sqrt{h^2}$ of 0.61 was reported for BFT with ridge regression BLUP (RR-BLUP) and of 0.56 with Bayesian LASSO [39]. A similar value of 0.55 was reported for Danish Duroc pigs [41]. Zhang et al. [9] reported an $r(GEBV_p, y^*)/\sqrt{h^2}$ of 0.38 for DFI in a Duroc population using a 80K SNP chip and the GBLUP method in a design with 1167 training animals and 196 validation animals. They reported a higher accuracy (0.45) when using a 650k SNP chip and the BayesB method. Prediction accuracies of GEBV for FCR and RFI are rarely reported

in the literature. Christensen et al. [8] reported a prediction accuracy of 0.16 for FCR using a bivariate ssGBLUP model. Jiao et al. [42] obtained a low prediction accuracy of 0.09 for RFI (measured as $r(GEBV_p, y^*)/\sqrt{h^2}$) using the BayesA method with 1047 training animals and 516 validation animals for the Duroc boars. Thus, overall in pig studies, prediction accuracies are low to moderate for ADG and BFT, and low for feed efficiency traits.

Prediction accuracies depending on the training set composition

Compared to FCR and RFI, ADG, BFT, and DFI showed different changes in prediction accuracy compared to scenario 1 when the structure of the training set was changed. For ADG, BFT, and DFI, removing the earlier generations of the target line from the training set (from scenarios 2 and 3 to scenarios 4, 5 and 6) generally decreased prediction accuracy to a lesser extent. The average and maximum relationships between the validation set and the training subsets were higher in scenarios 4, 5, and 6 than in scenario 1. The maximum relationship between the validation set and the training subsets, which was previously recommended as an indicator of potential accuracies [44], was lowest in scenario 1 and highest in scenario 4, likely due to changes in allele frequencies between the early and late generations within a line. This implies that the general decrease in accuracy in the scenarios 4, 5, and 6 could be attributed neither to these changes in relationships between sets, nor to the differences in prediction accuracies between lines. Moreover, the accuracy of GEBV resulting from ssGBLUP analyses should be less sensitive to the structure of the set of genotyped animals, and accordingly, to the strength of relationships between and within training and validation sets [45] because the **H** matrix aggregates information from both **A** and $\mathbf{A}_{22}$. This structure of the **H** matrix has two major effects on the GEBV of a given animal: first, it contributes the parent average EBV of the animal using the **A** matrix, and second, it adjusts for the different levels of relationships of the animal with other genotyped animals using the $\mathbf{A}_{22}$ matrix [45, 46]. de Roos et al. [19] reported that the benefits of combining populations in a training set are greatest when the populations have diverged for only a few generations and when the heritability of the trait is low. They also showed that increasing the number of animals from a given population in the training set increased prediction accuracy in that population. Considering that de Roos et al. [19] did not include the effect of selection in their simulations, this could partly explain our results for ADG, BFT, and DFI.

Impact of selection on accuracy and bias of predictions

The changes of accuracy across the scenarios were more diverse for RFI and FCR, with either increases or relatively similar accuracies compared to scenario 1. In some cases, the accuracy even increased as genotypes of closer generations were eliminated from the training set, which could be regarded as an effect of the different relationships between training and validation sets in these scenarios. Regarding the low prediction accuracy reported for FCR and RFI in our results and in the literature, denser SNP genotyping could probably increase the accuracy of predictions by better capturing the differences in LD between the lines. In addition, for low heritability traits, such as RFI in our study, large training populations have been reported to increase the accuracy of GEBV [47–49]. However, given that scenarios 5 and 6 resulted in accuracies that were comparable to that of the control scenario for FCR and in greater accuracies for RFI, they can be considered as optimum scenarios for an across-line genomic prediction program. Based on results from simulation, Pszczola et al. [50] declared that minimizing relationships within the reference population and maximizing them between training and validation sets maximizes the accuracy of genomic predictions. This means that including a diverse set of animals in the training set is desirable to some extent. This is consistent with our results for FCR and RFI, for which selection created two diverse sets of animals. For example, in scenario 6, including animals from G4 to G6 of the target line in the training set provided sufficient genetic links between training and validation sets, and animals from the G9 generation of the other line provided additional diversity to the training set. Overall, it seems that including animals from later generations of both lines (more diverse animals) in the training set contributed to higher accuracies of GEBV in the validation set for FCR and RFI. This might be because the SNP effects segregating in the validation set were better estimated with such a training set.

Overall, the comparison of accuracies between scenarios 4 to 6 and scenario 1 did not show an obvious effect of the removal of data of earlier generations from the training dataset. In a study using six levels of truncated data of past generations, accuracies of GEBV of young genotyped pigs were very similar for various reproductive traits [51].

Bias of genomic predictions

Our results showed that GEBV were more biased for traits that were more affected by selection, especially when early generations of the target line were not included in the training set. The scenarios that yielded better accuracies were not those with the smaller biases, except for FCR and RFI, for which predictions were low and their regression coefficients were systematically

below 1. The average and maximum relationships between training and validation sets did not affect the prediction biases in the same way for all traits, which could be due to the effect of selection. Selection in historical generations has been shown to result in considerable biases in EBV or GEBV [34, 52]. Tonussi et al. [53] emphasized that, to have accurate and unbiased GEBV with the ssGBLUP method, the **G** matrix should be compatible with the $\mathbf{A}_{22}$. Inappropriate merging of these matrices can originate from ignoring inbreeding in the structure of **A** and from changes in allele frequencies at QTL for the traits under selection. In our scenarios, the effect of selection in the last three generations of the validation sets was not explicitly accounted for. However, changes in marker allele frequencies in those generations were accounted for through the **G** matrix. Furthermore, the (co)variances used for genomic predictions were obtained from bivariate analyses including the selection criterion using the whole dataset (including validation generations). Therefore, there should be no effect of selection on the estimations of the variance components, and the prediction bias of the GEBVs should not be due to biased variance components. Computing separate accuracies and biases for sires (heavily selected) versus dams (not directly selected), could enable quantification of the effect of selection on the prediction biases. However, on the one hand, the dams had lower individual accuracies (no own phenotype), and on the other hand, only 18 sires were selected per line in these generations. Therefore, the resulting prediction accuracies and biases differed between sires and dams due to factors other than just the effect of selection and no clear conclusion could be reached. Finally, it should be mentioned that these three generations were combined into the validation set in our study to have enough individuals, but in practice, new candidates to be predicted pertain to a single unselected cohort, therefore this selection effect would be small and likely negligible.

Heritability, marker density and size of the training population have been shown to be important factors to control biases of prediction [54]. Therefore, the biases for some scenarios in this study could be explained by the low to medium heritability of the traits, the medium marker density information, and the small number of individuals in the training population. Testing similar prediction scenarios while ignoring pedigree relationships in the non-genotyped generations would lead to substantially biased predictions, especially for traits affected by selection (for instance, 1.61 for RFI predictions in the HRFI line for scenario 6). Combining full pedigree and genomic information appeared to limit bias, which is consistent with Tonussi et al. [53].

Conclusions

The results of our study show that genomic prediction using a training set that includes animals from related lines selected in different directions could be as accurate as genomic prediction using a within-line training set. Thus, this can be a solution to create a reference set in the case of small populations, or when ancestral samples are not available at low additional costs. Combined reference sets had better prediction accuracies for traits that were highly affected by selection, which can be attributed to the inclusion of more diverse animals in the training set. Overall, among all evaluated scenarios, scenarios 5 and 6 showed optimal accuracies in most cases, which is consistent with our hypothesis that data from a related line can be used in a combined training population for genomic predictions without losing prediction accuracy. Our results also proved that absence of phenotypic records from past generations did not affect prediction accuracy but increased bias of predictions. Some of these issues could be solved by using bivariate analyses or models with heterogeneous variances to better account for changes in variances with selection in different lines. Taken together, the results of our study provide insights into the design of reference populations for small populations, particularly when lines are being developed simultaneously, which is common in poultry and pig industries, and some plant breeding plans. This strategy can be recommended to initiate a genomic selection program when historical samples are not available, or when two lines are considered and genotyping costs need to be limited.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12711-020-00576-0.

Additional file 1. Figure S1. Correlation between $\mathrm{GEBV_p}$ and $\mathbf{y}^*$ and their SE as bars for the HRFI (a) and LRFI (b) lines. No scenario resulted in correlations that differed from those with scenario 1 based on a Williams t-test at 5%. RFI residual feed intake, ADG average daily gain, FCR feed conversion ratio, DFI daily feed intake, BFT backfat thickness. **Figure S2.** Correlation between $\mathrm{GEBV_p}$ and $\mathbf{y}^*$ divided by the square root of the heritability of corresponding traits for the HRFI (a) and LRFI (b) lines. RFI residual feed intake, ADG average daily gain, FCR feed conversion ratio, DFI daily feed intake, BFT backfat thickness.

Acknowledgements

The authors thank Andres Legarra for his comments and help. This study was financially supported by the French National Research Agency via the Micro-Feed project, under grant ANR-16-CE20-0003. The authors would like to thank the experimental farm staff for data collection and breeding of the animals. The authors are also grateful to Yvette Steyn for proof-reading the manuscript.

Authors' contributions

AA performed the statistical analyses and wrote the first draft of the paper. ED and YL performed the imputation and quality control of the genotypic data. AA, JR and HG participated in the design of the study. HG provided scientific supervision. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 23 September 2019 Accepted: 21 September 2020
Published online: 07 October 2020

References

1. Patience JF, Rossoni-Serão MC, Gutiérrez NA. A review of feed efficiency in swine: biology and application. J Anim Sci Biotechnol. 2015;6:33.
2. Gaines AM, Peterson BA, Mendoza OF. Herd management factors that influence whole herd feed efficiency. In: Patience JF, editor. Feed efficiency in swine. Wageningen: Wageningen Academic Publishers; 2012. p. 15–39.
3. Koch RM, Swiger LA, Chambers D, Gregory KE. Efficiency of feed use in Beef cattle. J Anim Sci. 1963;22:486–94.
4. Hoque MA, Suzuki K, Kadowaki H, Shibata T, Oikawa T. Genetic parameters for feed efficiency traits and their relationships with growth and carcass traits in Duroc pigs. J Anim Breed Genet. 2007;124:108–16.
5. Ollivier L, Gueblez R, Webb AJ, van der Steen HAM. Breeding goals for nationally and internationally operating pig breeding organisations. In: Proceedings of the 4th World Congress on Genetics applied to Livestock Production: 23–27 July 1990. Edinburgh; 1990
6. Pym RAE, Nicholls PJ. Selection for food conversion in broilers: direct and correlated responses to selection for body-weight gain, food consumption and food conversion ratio. Br Poult Sci. 1979;20:73–86.
7. Rexroad C, Vallet J, Matukumalli LK, Reecy J, Bickhart D, Blackburn H, et al. Genome to phenome: improving animal health, production and well-being a new USDA blueprint for animal genome research 2018–2027. Front Genet. 2019;10:327.
8. Christensen OF, Madsen P, Nielsen B, Ostersen T, Su G. Single-step methods for genomic evaluation in pigs. Animal. 2012;6:1565–71.
9. Zhang C, Kemp RA, Stothard P, Wang Z, Boddicker N, Krivushin K, et al. Genomic evaluation of feed efficiency component traits in Duroc pigs using 80 K, 650 K and whole-genome sequence variants. Genet Sel Evol. 2018;50:14.
10. Legarra A, Aguilar I, Misztal I. A relationship matrix including full pedigree and genomic information. J Dairy Sci. 2009;92:4656–63.
11. Misztal I, Legarra A, Aguilar I. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J Dairy Sci. 2009;92:4648–55.
12. Aguilar I, Misztal I, Johnson D, Legarra A, Tsuruta S, Lawlor T. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J Dairy Sci. 2010;93:743–52.
13. Christensen OF, Lund MS. Genomic prediction when some animals are not genotyped. Genet Sel Evol. 2010;42:2.
14. VanRaden PM, O'Connell JR, Wiggans GR, Weigel KA. Genomic evaluations with many more genotypes. Genet Sel Evol. 2011;43:10.
15. Carillier C, Larroque H, Robert-Granie C. Comparison of joint versus purebred genomic evaluation in the French multi-breed dairy goat population. Genet Sel Evol. 2014;46:67.
16. Lund MS, Su G, Janss L, Guldbrandtsen B, Brøndum RF. Genomic evaluation of cattle in a multi-breed context. Livest Sci. 2014;166:101–10.
17. Hayes BJ, Bowman PJ, Chamberlain AC, Verbyla K, Goddard ME. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet Sel Evol. 2009;41:51.
18. Olson KM, VanRaden PM, Tooker ME. Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. J Dairy Sci. 2012;95:5378–83.
19. de Roos AP, Hayes BJ, Goddard ME. Reliability of genomic predictions across multiple populations. Genetics. 2009;183:1545–53.
20. Zhang S-Y, Olasege BS, Liu D-Y, Wang Q-S, Pan Y-C, Ma P-P. The genetic connectedness calculated from genomic information and its effect on the accuracy of genomic prediction. PLoS One. 2018;13:e0201400.
21. Fangmann A, Bergfelder-Drüing S, Tholen E, Simianer H, Erbe M. Can multi-subpopulation reference sets improve the genomic predictive ability for pigs? J Anim Sci. 2015;93:5618–30.
22. Gilbert H, Billon Y, Brossard L, Faure J, Gatellier P, Gondret F, et al. Review: divergent selection for residual feed intake in the growing pig. Animal. 2017;11:1427–39.
23. Gilbert H, Bidanel J-P, Gruand J, Caritez J-C, Billon Y, Guillouet P, et al. Genetic parameters for residual feed intake in growing pigs, with emphasis on genetic relationships with carcass and meat quality traits. J Anim Sci. 2007;85:3182–8.
24. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.
25. Sargolzaei M, Chesnais JP, Schenkel FS. A new approach for efficient genotype imputation using information from relatives. BMC Genomics. 2014;15:478.
26. Careau V, Wolak ME, Carter PA, Garland T Jr. Limits to behavioral evolution: the quantitative genetics of a complex trait under directional selection. Evolution. 2013;67:3102–19.
27. Hadfield JD, Wilson AJ, Garant D, Sheldon BC, Kruuk LE. The misuse of BLUP in ecology and evolution. Am Nat. 2010;175:116–25.
28. Fernando R, Gianola D. Statistical inferences in populations undergoing selection or non-random mating. In: Gianola D, Hammond K, editors. Advances in statistical methods for genetic improvement of livestock. Berlin: Springer; 1990. p. 437–53.
29. Henderson C. Accounting for selection and mating biases in genetic evaluations. In: Gianola D, Hammond K, editors. Advances in statistical methods for genetic improvement of livestock. Berlin: Springer; 1990. p. 413–36.
30. Sorensen D, Fernando R, Gianola D. Inferring the trajectory of genetic variance in the course of artificial selection. Genet Res. 2001;77:83–94.
31. Misztal I, Tsuruta S, Lourenco D, Masuda Y, Aguilar I, Legarra A, et al. Manual for BLUPF90 family of programs. Athens: University of Giorgia; 2018.
32. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23.
33. Amadeu RR, Cellon C, Olmstead JW, Garcia AA, Resende MF Jr, Muñoz PR. AGHmatrix: R package to construct relationship matrices for autotetraploid and diploid species: a blueberry example. Plant Genome. 2016. https://doi.org/10.3835/plantgenome2016.01.0009.
34. Legarra A, Reverter A. Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. Genet Sel Evol. 2018;50:53.
35. Gunia M, Saintilan R, Venot E, Hoze C, Fouilloux MN, Phocas F. Genomic prediction in French Charolais beef cattle using high-density single nucleotide polymorphism markers. J Anim Sci. 2014;92:3258–69.
36. Steiger JH. Tests for comparing elements of a correlation matrix. Psychol Bull. 1980;87:245–51.
37. Williams EJ. The comparison of regression variables. J R Stat Soc Series B Stat Methodol. 1959;21:396–9.
38. Revelle WR. Package Psych V1.8.12: Procedures for psychological, psychometric, and personality research. Evanston: Northwestern University; 2019.
39. de Campos CF, Lopes MS, Silva FF, Veroneze R, Knol EF, Lopes PS, et al. Genomic selection for boar taint compounds and carcass traits in a commercial pig population. Livest Sci. 2015;174:10–7.
40. Do DN, Janss LL, Strathe AB, Jensen J, Kadarmideen H, editors. Genomic prediction and genomic variance partitioning of daily and residual feed intake in pigs using Bayesian Power Lasso models. In: Proceedings of the 10th World Congress on Genetics Applied to Livestock Production: 17–22 Aug 2014; Vancouver, 2014.
41. Guo X, Christensen OF, Ostersen T, Wang Y, Lund MS, Su G. Genomic prediction using models with dominance and imprinting effects for backfat thickness and average daily gain in Danish Duroc pigs. Genet Sel Evol. 2016;48:67.
42. Jiao S, Maltecca C, Gray KA, Cassady JP. Feed intake, average daily gain, feed efficiency, and real-time ultrasound traits in Duroc pigs: I. Genetic parameter estimation and accuracy of genomic prediction. J Anim Sci. 2014;92:2377–86.
43. Reverter A, Tier B, Johnston DJ, Graser HU. Assessing the efficiency of multiplicative mixed model equations to account for heterogeneous variance across herds in carcass scan traits from beef cattle. J Anim Sci. 1997;75:1477–85.

44. Clark SA, Hickey JM, Daetwyler HD, van der Werf JH. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet Sel Evol. 2012;44:4.
45. Lourenco DA, Fragomeni BO, Tsuruta S, Aguilar I, Zumbach B, Hawken RJ, et al. Accuracy of estimated breeding values with genomic information on males, females, or both: an example on broiler chicken. Genet Sel Evol. 2015;47:56.
46. Misztal I, Tsuruta S, Aguilar I, Legarra A, VanRaden P, Lawlor T. Methods to approximate reliabilities in single-step genomic evaluation. J Dairy Sci. 2013;96:647–54.
47. Goddard M. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica. 2009;136:245–57.
48. Hayes BJ, Bowman PJ, Chamberlain A, Goddard M. Invited review: genomic selection in dairy cattle: Progress and challenges. J Dairy Sci. 2009;92:433–43.
49. Hoze C, Fritz S, Phocas F, Boichard D, Ducrocq V, Croiseau P. Efficiency of multi-breed genomic selection for dairy cattle breeds with different sizes of reference population. J Dairy Sci. 2014;97:3918–29.
50. Pszczola M, Strabel T, Mulder HA, Calus MP. Reliability of direct genomic values for animals with different relationships within and to the reference population. J Dairy Sci. 2012;95:389–400.
51. Lourenco DA, Misztal I, Tsuruta S, Aguilar I, Lawlor TJ, Forni S, et al. Are evaluations on young genotyped animals benefiting from the past generations? J Dairy Sci. 2014;97:3930–42.
52. Bijma P. Accuracies of estimated breeding values from ordinary genetic evaluations do not reflect the correlation between true and estimated breeding values in selected populations. J Anim Breed Genet. 2012;129:345–58.
53. Tonussi RL, de Oliveira Silva RM, Magalhães AFB, Espigolan R, Peripolli E, Olivieri BF, et al. Application of single step genomic BLUP under different uncertain paternity scenarios using simulated data. PLoS One. 2017;12:e0181752.
54. Karimi K, Sargolzaei M, Plastow GS, Wang Z, Miar Y. Opportunities for genomic selection in American mink: a simulation study. PLoS One. 2019;14:e0213873.

Publisher's Note

Annexe 4

Delpuech *et al. Genet Sel Evol* (2021) 53:49
https://doi.org/10.1186/s12711-021-00642-1

GSE Genetics Selection Evolution

RESEARCH ARTICLE Open Access

Identification of genomic regions affecting production traits in pigs divergently selected for feed efficiency

Emilie Delpuech[1], Amir Aliakbari[1], Yann Labrune[1], Katia Fève[1], Yvon Billon[2], Hélène Gilbert[1] and Juliette Riquet[1*]

Abstract

Background: Feed efficiency is a major driver of the sustainability of pig production systems. Understanding the biological mechanisms that underlie these agronomic traits is an important issue for environment questions and farms' economy. This study aimed at identifying genomic regions that affect residual feed intake (RFI) and other production traits in two pig lines divergently selected for RFI during nine generations (LRFI, low RFI; HRFI, high RFI).

Results: We built a whole dataset of 570,447 single nucleotide polymorphisms (SNPs) in 2426 pigs with records for 24 production traits after both imputation and prediction of genotypes using pedigree information. Genome-wide association studies (GWAS) were performed including both lines (global-GWAS) or each line independently (LRFI-GWAS and HRFI-GWAS). Forty-five chromosomal regions were detected in the global-GWAS, whereas 28 and 42 regions were detected in the HRFI-GWAS and LRFI-GWAS, respectively. Among these 45 regions, only 13 were shared between at least two analyses, and only one was common between the three GWAS but it affects different traits. Among the five quantitative trait loci (QTL) detected for RFI, two were close to QTL for meat quality traits and two pinpointed novel genomic regions that harbor candidate genes involved in cell proliferation and differentiation processes of gastrointestinal tissues or in lipid metabolism-related signaling pathways. In most cases, different QTL regions were detected between the three designs, which suggests a strong impact of the dataset structure on the detection power and could be due to the changes in allelic frequencies during the establishment of lines.

Conclusions: In addition to efficiently detecting known and new QTL regions for feed efficiency, the combination of GWAS carried out per line or simultaneously using all individuals highlighted chromosomal regions that affect production traits and presented significant changes in allelic frequencies across generations. Further analyses are needed to estimate whether these regions correspond to traces of selection or result from genetic drift.

Background

Feed efficiency is a major driver of the sustainability of pig production systems. It represents from 50 to 83% of production costs depending on countries and systems [1]. Feed efficiency is also a major lever to reduce the environmental footprints of production [2]. In pig production, the cost of feeding is usually measured by computing the feed conversion ratio (FCR). Indeed, FCR is a ratio between two traits of interest in most breeding schemes (feed intake and growth rate), and its incorporation in selection indexes makes it difficult to accurately anticipate responses to selection on this trait and the correlated traits [3]. In 1963, Koch et al. [4] proposed residual feed intake (RFI) as an alternative to quantify feed efficiency and overcome the limits of FCR. RFI is the

*Correspondence: juliette.riquet@inrae.fr
[1] GenPhySE, Université de Toulouse, INRAE, ENVT, 31320 Castanet-Tolosan, France
Full list of author information is available at the end of the article

difference between individual feed intakes and predicted feed intake for the animals' maintenance and production requirements. It is generally computed as a multiple linear regression of daily feed intake on production traits (growth rate and body composition traits in growing animals), and on the average metabolic body weight of the animal during the growth period, as an indicator of maintenance requirements. As a result, selection for RFI generates limited correlated responses on the other production traits, as shown in several selection experiments in pigs [5, 6], and other species [7]. However, recording accurately individual feed intake for pigs raised in groups is costly, and large efforts are devoted to facilitate the improvement of feed efficiency, by either identifying biomarkers [8, 9] or genomic markers (for instance [10, 11]). In spite of these efforts, the difficulty to identify quantitative trait loci (QTL) or genomic variants that affect feed efficiency related traits is illustrated by the PigQTLDB statistics [12]: only 394 QTL are listed for feed conversion traits, and 350 for feed intake traits, whereas more than 2000 are reported for growth traits, and more than 3200 for fatness traits (PigQTLDB, accessed September 2020, https://www.animalgenome.org/cgi-bin/QTLdb/SS/index). Genomic information acquired from established divergent lines for the trait of interest can be used to increase the power of detection of genomic variants for lowly heritable or highly polygenic traits, such as RFI in pigs [10] and litter traits in rabbits [13].

In this study, our aim was to identify genomic regions that affect RFI and other production traits in two pig lines that have been divergently selected for RFI during nine generations [5], by combining extensive genotyping of all breeding animals of the lines, and extensive phenotyping of their progeny. GWAS were applied to growth, feed intake and feed efficiency, carcass composition and meat quality traits on the full dataset. Different subsets of the population were used to be able to suggest biological hypotheses regarding the genetic background of the traits in the two divergent lines, and to decipher whether the chromosomal regions that affecting RFI differ between lines.

Methods

Ethic statement

All pigs were reared in compliance with national regulations and according to procedures approved by the French Veterinary Services at INRAE experimental facilities. The care and use of pigs were performed following the guidelines edited by the French Ministries of High Education, Research and Innovation, and of Agriculture and Food (http://ethique.ipbs.fr/sdv/charteexpeanimale.pdf).

Design

The data were obtained from a divergent selection experiment on RFI carried out at the INRAE experimental unit GenESI since 2000 (Surgères, France, https://doi.org/10.15454/1.5572415481185847E12), on growing pigs from the French Large-White (LW) population. Selection procedures were previously described by Gilbert et al. [5]. In brief, the lines were established from 30 matings of LW animals (F0). From these litters, 116 males were tested to select the six most efficient (LRFI) and six least efficient (HRFI) males as founders of two divergent lines, and about 40 pairs of sibs were randomly assigned to each line. In the following generations, from G1 to G9, 96 males from each line were tested for RFI to select six extreme low or high boars depending on the line. In addition, 35 to 40 females were randomly chosen within-line in each generation to produce the next generation. No selection was applied for females. After nine generations of selection, an average inbreeding of 19% was estimated in the lines. From G1, matings were organized for at least two successive litters. Until G5, the first litter provided boar candidates for selection and future breeding females, and castrated males and females from the second parity were tested to evaluate the direct and correlated responses to selection on major production traits, including carcass composition and meat quality traits. In generation G9, the responses to selection reached −165 g/day (LRFI line–HRFI line) for RFI (3.84 genetic standard deviations (σ_g)), and −270 g/day for DFI (2.11 σ_g) (Table 1). After G5, selection was applied to parity 4 or 5, and responses to selection were measured on pigs born in parities 2 and 3. Hereafter, the breeding animals are called "breeders" and animals tested for responses to selection are called "response animals".

Phenotypes

For this study, 2426 phenotyped response animals were used, which corresponds to about 48 females and 48 castrated males per line in each generation G1 to G5, plus 700 response animals per line distributed in generations G6 to G9. All animals were raised during the growing-finishing period (~28 kg to ~107 kg) in the same growing-finishing unit comprising four rooms of four pens, each equipped with a single-place electronic feeder (ACEMA 64; Skiold Acemo, Pontivy, France). Each animal had records for body weight (BW0 at the start of the test and BW1 before slaughter) and daily feed intake (DFI) to compute average daily gain (ADG) and feed conversion ratio (FCR) during the test period. The dressing percentage (DP) was computed based on weight records of warm carcass at slaughter. Twenty four hours after slaughter, backfat thickness

Table 1 Number of QTL identified for each trait with the three groups of association studies

Trait	h^2	Genetic differences in G9 (σ_g)	Global	HRFI	LRFI	Total
DFI	0.41	2.11	2	1	3	6
ADG	0.5	0.15	1	*3*		4
FCR	0.42	2.46	2		2	4
RFI	0.13	3.84	3		2	5
carcBFT	0.4	0.037	4	*1*	*4*	9
a*_GM	0.29	0.38	2		1	3
a*_GS	0.26	0.12		4	4	8
b*_GM	0.24	0.09	1		1	2
b*_GS	0.32	1.14	6	*4*	*1*	11
L*_GM	0.2	0.38	1	*5*		6
L*_GS	0.33	2.12	2	3	4	9
pH24h_AD	0.41	1.39	2		*3*	5
pH24h_GS	0.39	1.98	4	1	1	6
pH24h_LM	0.32	1.45	4	3	4	11
pH24h_SM	0.38	1.74	3	1	1	5
WHC	0.04	0.68	3	*5*		8
MQI	0.33	1.92	4	1	1	6
LMCcalc	0.59	1.31	3		1	4
DP	0.36	0.93	3	*1*	*6*	10
Belly_W	0.28	1.90			2	2
BF_W	0.43	0.9	2	1	3	6
Ham_W	0.51	0.97	2	1	1	4
Loin_W	0.54	1.69	2		1	3
Shoulder_W	0.38	1.11		1	1	2
Total			56	36	47	139

Association studies on the full population (global-GWAS, *Global*) and for each line separately (HRFI-GWAS, *HRFI* and LRFI-GWAS, *LRFI*) were performed. Traits with more than three different QTL between the HRFI-GWAS and LRFI-GWAS analyses are indicated in italic characters. For each trait h^2 = heritability, and responses to selection expressed in genetic standard deviations of the trait are reported as computed by Gilbert et al. [27]

DFI: daily feed intake; ADG: average daily gain; FCR: feed conversion ratio; RFI: residual feed intake; *carcBFT*: backfat thickness measured on carcass; a*_GM: a* measured on the *gluteus medius* muscle; a*_GS: a* measured on the *gluteus superficialis* muscle; b*_GM: b* measured on the *gluteus medius* muscle; b*_GS: b* measured on the *gluteus superficialis* muscle; L*_GM: L* measured on the *gluteus medius* muscle; L*_GS: L* measured on the *gluteus superficialis* muscle; *pH24h_AD*: pH 24 h after slaughter measured on the adductor femoris muscle; pH24h_GS: pH 24 h after slaughter measured on the *gluteus superficialis* muscle; pH24h_LM: pH 24 h after slaughter measured on the *longissimus dorsi* muscle; pH24h_SM: pH 24 h after slaughter measured on the *semimembranosus* muscle; WHC: water holding capacity of the *gluteus superficialis* muscle; MQI: meat quality index; LMCcalc: lean meat content of the carcass; DP: carcass dressing percentage; Belly_W: belly weight; BF_W: backfat weight; Ham_W: ham weight; Loin_W: loin weight; Shoulder_W: shoulder weight

measured on carcass (carcBFT), and the weights of ham (Ham_W), loin (Loin_W), belly (Belly_W), shoulder (Shoulder_W), and backfat (BF_W), following a standardized cut, were recorded on the cold half carcass. The lean meat content (LMCcalc) was estimated from a linear combination of the weights of carcass ham, loin, and backfat, expressed as a percentage of the half-carcass weight [14]: LMC (%) = 25.08 − 1.23 backfat (%) + 0.87 loin (%) + 0.73 ham (%). Meat quality measurements included pH on the *adductor femoris* (AD), *semimembranosus* (SM), *gluteus superficialis* (GS), and *longissimus dorsi* muscles (LM), colorimetry L*, a* and b* on GS and *gluteus medius* muscle (GM), and water-holding capacity (WHC) assessed on GS according to the procedure described by Charpentier et al. [15]. Finally, a meat quality index (MQI) was calculated from measurements of the pH on SM, L* on GS and WHC according to the model proposed by Tribout et al. [16]. RFI was defined as the residual of a multiple linear regression as follows: $RFI = DFI - (1.48 \times ADG) + (23.2 \times LMCcalc) - (99.1 \times AMBW)$, where AMBW is the average metabolic body weight during the test period and is equal to $(BW1^{1.6} - BW0^{1.6})/[1.6\ (BW1 - BW0)]$ [17]. Contemporary group (group of around 45 animals born in the same week and contemporarily tested in a given room), gender and pen size were added as fixed effects in the model, as described by Gilbert et al. [5].

Genotyping

Genomic DNA was purified from individual biological samples of the sires and dams of all generations using standard protocols. Over the time of the study, two different Illumina medium-density SNP chips were used according to the genotyping protocols defined by the supplier (Technological Center, Genomics and Transcriptomics Platform, CRCT Toulouse). A first genotyping batch comprising 286 animals (12 sires from each generation G0 to G6, and G0, G3 and G6 dams) was genotyped for 64,232 SNPs using the Porcine SNP60v2 BeadChip (60K SNPs chip), and a second batch of 1356 animals (complementary breeding animals of the generations G0 to G6 and sires and dams of the following generations) was genotyped using the Porcine HD Array GGP chip comprising 68,516 SNPs (70K SNPs chip). Genotypes were obtained using the Genome Studio software (V2.0.4) and coded as 0, 1 and 2 corresponding, respectively, to individuals homozygous for the minor allele, heterozygous and homozygous for the major allele. In addition, 32 G0 founders equally distributed between the lines (12 G0 sires, and 20 G0 dams that contributed most to the subsequent generations based on pedigree information) were genotyped with the Affymetrix Axiom Porcine HD Genotyping Array chip (Gentyane Platform, UMR 1095 INRAE Clermont-Ferrand) consisting of 658,692 SNPs (650K SNPs chip).

For each SNP panel, quality control was performed using the PLINK software (V1.90) [18]: SNPs with a call frequency (CF) lower than 95% and a minor allele frequency (MAF) lower than 1% were excluded, and animals with a call rate (CR) lower than 90% were discarded. Deviations from Hardy–Weinberg equilibrium were also assessed with a p-value of 10^{-10}. Unmapped SNPs and SNPs located on the sex chromosomes were removed based on the Sscrofa11.1 assembly of the reference genome (https://www.ensembl.org/Sus_scrofa/Info/Index) [19].

Imputation of genotypes

Two successive imputations were performed using the FImpute software [20]. A first level of imputation was performed with markers on the 60K and 70K SNPs chips, based on 29,957 common SNPs, to homogenize the medium-density genotyping data available for the 1632 breeders of the lines. This leads to an intermediate dataset of 66,988 SNPs that are imputed from both medium-density (MD) chips (60K and 70K SNPs chips). In a second step, the genotypes of the high-density (HD) SNPs chip were imputed for all breeders using the HD SNP genotypes of the 32 G0 founders. A set of common 45,708 SNPs was available between the MD imputed genotypes and the HD SNP chip. Finally, 570,447 SNPs distributed along the 18 pig autosomes were available for 1632 breeding animals.

To evaluate imputation accuracy, first, five successive batches of 1000 SNPs were randomly selected among the common SNPs between the 60 and 70K SNP chips. For each SNP batch, the genotypes of these SNPs were set as missing for all animals genotyped with the 60K SNPs chip and imputed from the 70K SNPs chip information. Therefore, 5000 SNPs with real and imputed genotypes were used to compute Pearson's correlations for each of the 286 pigs with 60K genotypes. Similarly, five batches of 1000 SNPs were randomly selected among the common SNPs between both MD SNP chips, animals genotyped with the 70K SNPs array were re-coded as missing, and Pearson's correlations between true and imputed genotypes were computed for the 1346 animals with 70K SNP genotypes. Then, to evaluate the imputation quality to the HD level, the same strategy of removing successively five batches of 1000 SNPs from the data was applied using SNPs that were in common among the three chips. In addition, a leave-one-out approach was applied to the 32 individuals with HD genotypes to evaluate the imputation accuracy.

In addition, a multi-dimensional scaling (MDS) analysis was performed using the *cmdscale()* function in the R software (V.3.6.2, R Core Team 2019) based on a identity-by-state matrix constructed with the PLINK software [18].

Predicted genotypes in response animals

Response animals did not have genotypes themselves. An average expected genotype was computed for each animal from the imputed 650K genotypes of their parents. For each SNP, each individual was given the average genotype of the parents (0, 0.5, 1, 1.5 or 2), thus within a litter, all animals were assigned the same genotypes. Thus, depending on the class of genotypes, the obtained genotype represented an approximation of the real genotype: (i) genotypes 0 and 2 were certain, as they resulted from two homozygous parents for the same allele (0×0 0 and 2×2 2), (ii) genotypes 0.5 and 1.5 included combinations of a homozygous genotype for one allele and a heterozygous genotype (0×1 0 or 1 and 1×2 1 or 2), and (iii) genotype 1 was the most heterogeneous class, with a mixture of true genotypes (0×2 1) and uncertain genotypes (1×1 0 or 1 or 2). Animals whose parents had a missing genotype were excluded from the analysis.

Genome-wide association studies

GWAS analyses were performed using the GEMMA software (version 0.97) [21] on response animals with their own phenotypes and their average genotypes from the parents. Phenotypes were adjusted for significant fixed

effects and covariates (pen size, herd, sex, and contemporary groups for in vivo measurements, slaughter date as fixed effects, and slaughter age as covariate for traits recorded at the abattoir, and slaughter BW as covariate for carcBFT) using linear models as proposed in Aliakbari et al. [22]. The resulting residues were integrated as phenotypes in GEMMA. To account for the structure of the population in the GWAS analyses, a pedigree relationship matrix **A** was computed. Association analyses were performed on the 24 traits available for the 2426 response animals.

The statistical model used to test one marker at a time was $\mathbf{y} = \mathbf{x}\beta + \mathbf{Zu} + \boldsymbol{\varepsilon}$, where $\mathbf{y}$ is the vector of adjusted phenotypes for all individuals; $\mathbf{x}$ is a vector of genotypes at the tested marker; β is the effect of the tested marker; $\mathbf{u}$ is a vector of random additive genetic effects distributed according to $N(0, \mathbf{A}\lambda\tau^{-1})$, with λ the ratio of the additive genetic variance and the residual variance τ^{-1} and $\mathbf{Z}$ the incidence matrix (identity matrix in this case); $\boldsymbol{\varepsilon}$ is a vector of residuals $N(0, \mathbf{I}\tau^{-1})$, with $\mathbf{I}$ the identity matrix. In GEMMA, an efficient exact algorithm is implemented to first estimate λ, and next derive $\widehat{\beta}$ and $\widehat{\tau}$ for each marker [23].

Three types of populations were considered for GWAS. First, the full dataset, which combines the two lines, was analyzed in a global analysis (thereafter called global-GWAS). Then, to evaluate if some QTL were segregating in one line only, the analyses were repeated within line (thereafter called lines-GWAS, or HRFI-GWAS and LRFI-GWAS when only one line was referred to).

For each analysis, the distributions of the test statistics of the GWAS of each trait were checked using quantile–quantile plots (Q-Q plot), and we computed the regression coefficients of the observed to the expected distribution under H_0. Inflation factors were on average 1.17 ± 0.15 for all analyses, indicating low deviations from the distribution of the test statistic under H_0. However, a correction factor was applied to all the analyses to control type-I errors, by dividing each chi square statistic by the corresponding inflation factor, following the genomic control approach proposed by Devlin and Roeder [24]. The test nominal p-values were computed according to this new chi square statistic.

To account for the multiple testing issue in the computation of genome-wide type-I errors, the significance threshold was obtained after a Bonferroni correction as follows:

$$-\log 10\left(\frac{0.05}{\sum_{i=1}^{nbchr} number\ of\ independent\ tests_i)}\right),$$

where the number of independent tests was computed as the sum of the number of independent tests for each chromosome. For each chromosome, this number was the number of principal components required to describe 99.6% of the genotype variability, obtained from a principal component analysis applied to the correlation matrix between genotypes of the SNPs on the considered chromosome (square root (r^2) of linkage disequilibrium (LD) between each pair of SNPs, Gao et al. [25]). The resulting genome-wide threshold (4.5 corresponding to 1690 independent tests) was used to select significant associations for each type of analysis. In addition, a cut-off of 3 (chromosome-wide threshold) was used only to assess whether a significant region identified in one analysis was suggestive in another one.

To define QTL intervals, the genome was divided into 1-Mb windows following the Sscrofa11.1 assembly of the reference genome. First, for each analysis (HRFI-GWAS, LRFI-GWAS and global-GWAS performed for each trait), the 1-Mb windows with at least one SNP with a significant p-value at 5% genome-wide ($-\log_{10}(p\text{-value}) \geq 4.5$) were retained, and adjacent windows with significant signals were combined into a single "QTL-window" per trait. In a second step, all the QTL windows were combined across traits using the same approach as above: adjacent and overlapping QTL-windows were fused, thus allowing the definition of a complete list of "QTL-regions". When a QTL-region was significant for several traits, for each one, the most significant marker and the associated allelic substitution effect was retained to tag the QTL (trait × region) for this trait in further analyses – thereafter called SNP-QTL.

The QTL positions were compared to previously mapped QTL in pigs using the pigQTLdb database [12], and QTL significant for RFI trait were screened for functional candidate genes using the Ensembl annotation V.101 (August 2020).

Changes in allelic frequencies of SNP-QTL

The power of detection in GWAS is strongly influenced by the allelic frequencies of the analyzed markers [26]. Within each QTL region, the different SNP-QTL were considered to examine the changes in allele frequencies with line selection. It should be noted that in addition to selection, changes in allele frequencies can also be due to genetic drift, especially in small closed populations. For instance, under the Wright-Fisher model (panmixia, no selection, $N = 40$) in our lines, genetic drift would result in generation 9 in standard deviations of allele frequencies of 0.164 for SNPs with an initial frequency of 0.5. However, our objective was not to test if allele frequencies responded to selection, but to illustrate changes in allele frequencies with time, accounting for all generations, in QTL regions. These SNP-QTL allele frequencies were estimated for the response animal genotypes,

i.e. from their average genotypes. To investigate how selection affected allele frequencies, and thus power of detection, allele frequencies were computed by adding animals from one generation at a time, starting from G1 individuals only. Then, the allele frequencies by adding G2 response animals were obtained by combining genotypes of G1 and G2 response animals, and so on until G9. The estimated frequencies in G9 (using all the animals from G1 to G9) corresponded to the informativeness of the markers used in the main lines-GWAS. In each line, a regression of the generation number (1 to 9) on the SNP allele frequencies was then applied to test changes in allelic frequencies on cumulative datasets across generations. For each SNP-QTL, the significance of the slope was tested in each line using a Wald test. The comparison of the slopes (the regression coefficients of the allelic frequencies) between lines highlighted four distinct cases: (i) markers with frequencies that did not change with line selection (no slope differed from zero with the Wald tests), (ii) markers co-selected in the two lines (slopes differed from zero and had identical signs), (iii) markers selected in opposite directions in the lines (slopes differed from zero with different signs), and (iv) markers with frequencies that changed in one line only (slope different from zero in one line only). Using only the significant slope values, a QTL evolution score was computed for each SNP-QTL as (9 generations * (|slope $_{HRFI}$| +|slope $_{LRFI}$|)) and to summarize their evolution per trait, an average score over all SNP-QTL for each trait was computed.

Results

Quality control and imputation of genotypes

True SNP genotyping data were available for all sires and dams from G0 to G9. The quality control of the genotypes was carried out first for each SNP chip independently. With a CR threshold of 90%, 10 animals genotyped with the 70k SNP chip and no individual genotyped with the 60K and 650K SNP chips were discarded (see Additional file 1: Table S1). For the SNPs, 15,114 SNPs from the 60K SNP chip (5776 for CF < 95% and 9125 for MAF < 1%), 11,891 SNPs from the 70K SNP chip (5323 for CF < 95% and 6568 for MAF < 1%), and 99,587 SNPs from the HD SNP chip (53,735 for CF < 95% and 45,852 for MAF < 1%) were removed. No SNP was removed with the Hardy–Weinberg equilibrium filter. In total, genotypes of 286 animals for 49,118 SNPs from the 60k SNP chip, genotypes for 1346 animals for 56,625 SNPs from the 70K SNP chip, and finally genotypes for 32 animals for 559,105 SNPs from the HD SNP chip were retained for further analyses (see Additional file 2: Table S2).

To obtain HD genotypes for all parents in the design, two successive runs of imputations were performed. First, the imputation of the missing genotypes on each MD support (60K and 70K SNP chips) allowed us to obtain genotypes for 66,988 SNPs for all sires and dams. The imputation accuracy was on average 0.995 regardless of the generation of the imputed individuals (see Additional file 3: Figures S1a and 1b). A second run of imputation was applied to all breeding animals from the 32 founder individuals genotyped with the HD SNP chip. The imputation accuracy was also high, with average accuracies around 0.979 (see Additional file 3: Figure S1c). A few animals in G0 and G3 had accuracies lower than 0.97. The accuracy estimated via the leave-one-out approach confirmed the values estimated with the correlations, with an average of 0.975 (see Additional file 3: Figure S1d). In total, genotypes for 570,447 SNPs were obtained for all parents from G0 to G9.

An MDS analysis was performed on the genotype matrix to represent the changes in genomic content of the lines with generations (Fig. 1). The first component corresponded to the dispersion of individuals according to the lines, and the second component corresponded to the successive generations in both lines.

Genome-wide association studies

From the imputed genotypes of all parents, an average genotype was computed for all response animals. Thus, genotypes coded 0, 0.5, 1, 1.5 or 2 were available for 2426 individuals. In total, the design included 596 full-sib families including 4.07 (±2.9) individuals on average. Within a sibling, all individuals shared

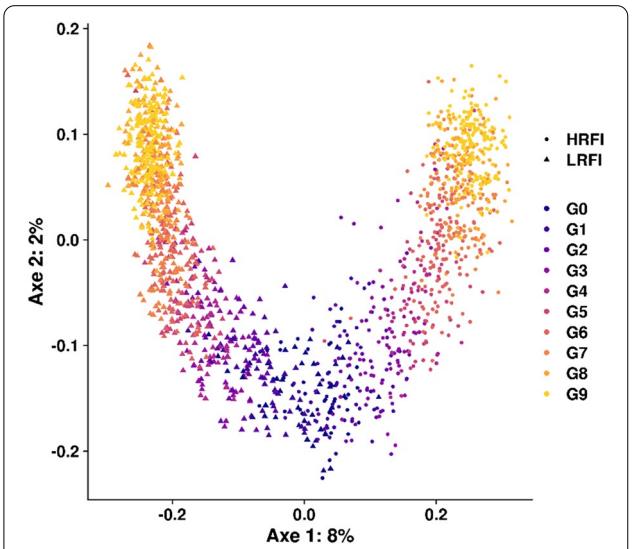

Fig. 1 The first two axes of the multidimensional scaling (MDS) analysis, based on the 570,447 genotypes. Points represent individuals (corresponding to all sires and dams of the population, N = 1632) and colors are generations

the same average genotype. The proportions of the five possible genotypes were estimated for each SNP and each individual in the design, with indication of their uncertainty. For each SNP, the proportion of certain genotypes, corresponding to classes 0, 1 (half of them) and 2, represented 1276 genotypes on average, i.e. 53% of the individuals, with a median of 1130 genotypes that are certain, this proportion being higher for SNPs with an extreme MAF. In addition, for each individual, among the 66,988 SNPs for the MD imputed genotypes considered in the calculation, from 31,132 to 40,852 SNPs (an average of 35,232 SNPs) were predicted with certainty (see Additional file 4: Figure S2).

First, association studies corresponding to global-GWAS were carried out on all response animals, for each of the 24 traits. Significant regions were selected by applying a genome-wide threshold of 4.5. Forty-five regions of 1 Mb (31 regions), 2 Mb (6 regions), 3 Mb (7 regions) or 8 Mb (1 region) were significant for at least one trait, corresponding to 56 QTL-windows (trait × region) for the global-GWAS. For all traits (except Belly_W, Shoulder_W and a*_GS), at least one QTL was detected in these analyses (Fig. 2), the list and characteristics of these QTL are reported in Additional file 5: Table S3.

To assess whether the identified QTL regions were identical and shared between the two lines, complementary GWAS analyses were performed per line, using either the set of individuals from G1 to G9 of the HRFI line or the set of individuals from G1 to G9 of the LRFI line. The QTL identified with the three analyses were compared (Table 1 and Fig. 2). As an example of the outcome of these analyses, Manhattan plots for RFI obtained with the global-GWAS and lines-GWAS are reported in Additional file 6: Figure S3. For the analyses performed by line, the number of regions detected

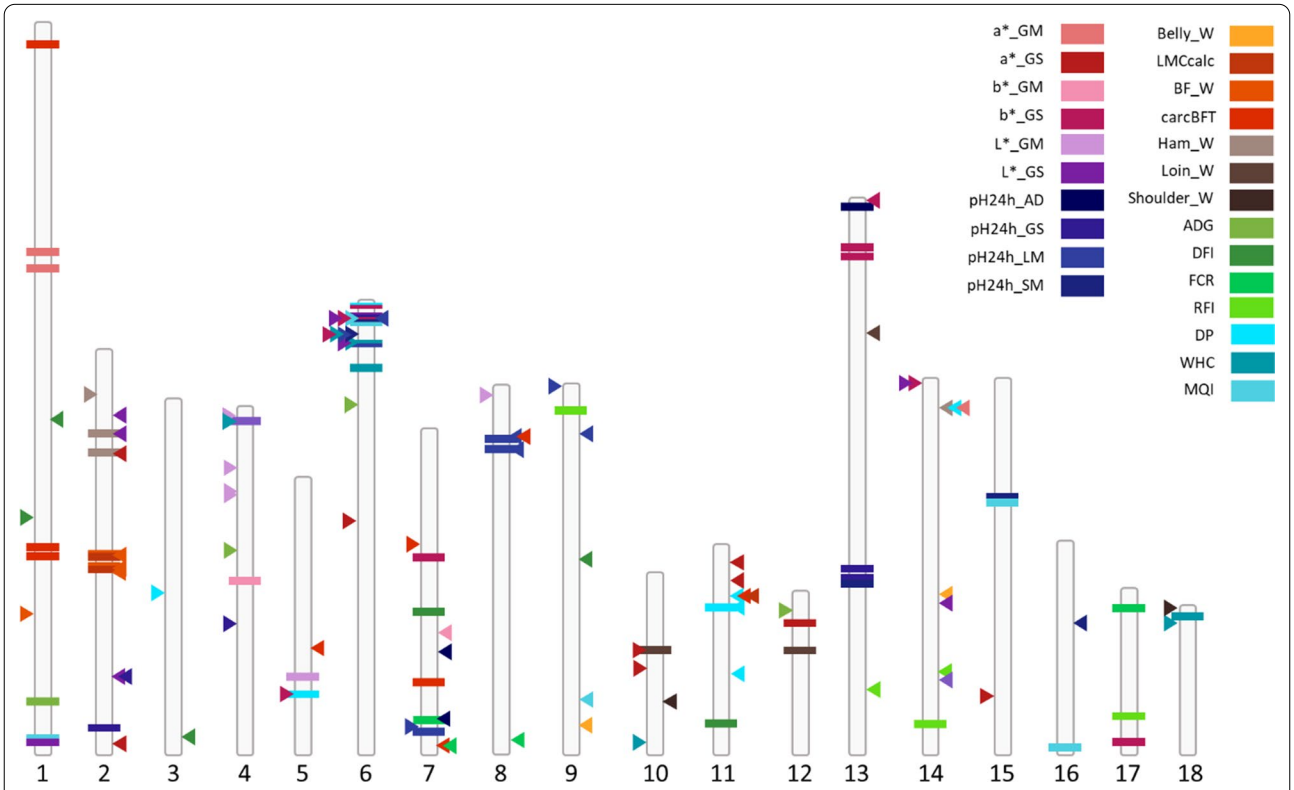

Fig. 2 Location of all SNP-QTL identified on the 18 autosomes from the Global-GWAS, LRFI-GWAS and HRFI-GWAS. The SNP-QTL corresponding to Global-GWAS are represented by horizontal bars, LRFI-GWAS by arrows to the right of the chromosomes and HRFI-GWAS by arrows to the left of the chromosomes. Each color represents one of the 24 traits; LRFI: low RFI line; HRFI: high RFI line; DFI: daily feed intake; ADG: average daily gain; FCR: feed conversion ratio; RFI: residual feed intake; carcBFT: backfat thickness measured on carcass; a*_GM: a* measured on the *gluteus medius* muscle; a*_GS: a* measured on the *gluteus superficialis* muscle; b*_GM: b* measured on the *gluteus medius* muscle; b*_GS: b* measured on the *gluteus superficialis* muscle; L*_GM: L* measured on the *gluteus medius* muscle; L*_GS: L* measured on the *gluteus superficialis* muscle; pH24h_AD: pH 24 h after slaughter measured on the adductor femoris muscle; pH24h_GS: pH 24 h after slaughter measured on the *gluteus superficialis* muscle; pH24h_LM: pH 24 h after slaughter measured on the *longissimus dorsi* muscle; pH24h_SM: pH 24 h after slaughter measured on the *semimembranosus* muscle; WHC: water holding capacity of the *gluteus superficialis* muscle; MQI: meat quality index; LMCcalc: lean meat content of the carcass; DP: carcass dressing percentage; Belly_W: belly weight; BF_W: backfat weight; Ham_W: ham weight; Loin_W: loin weight; Shoulder_W: shoulder weight

for a trait could differ between lines. For instance, more loci were detected in the HRFI line for ADG, b*_GS, L*_GM and WHC, whereas more regions were detected in the LRFI line for carcBFT, pH24h_AD and DP. In the HRFI line, 36 QTL were identified in 28 regions, and in the LRFI line, 47 QTL were identified in 42 regions. Only one region overlapped between the two lines: on SSC6, a region located between 7 to 10 Mb affected pH24h_LM in LRFI and L*_GS, b*_GS, and MQI in HRFI, which are highly correlated traits related to meat quality (Fig. 3).

Cut weights were the traits with the smallest number of QTL (1 to 3 per analysis) (Table 1). Meat quality measurements had the largest number of QTL (up to 6). Nineteen regions associated with growth, feed intake, and feed efficiency were detected, including five regions associated with RFI and four with FCR.

Thirteen regions were shared between the 45 regions identified in the global-GWAS and the 69 unique regions from the analyses per line, with only five common regions between the global-GWAS and HRFI-GWAS analyses, nine common regions between the global-GWAS and LRFI-GWAS, and the SSC6 region described above detected in the three analyses (Fig. 3a). Among these regions, only nine QTL (trait x region) were identified jointly in the global-GWAS and in one of the lines-GWAS (Fig. 3b), and none was shared in the three analyses. Thus, very few QTL were common between the three GWAS (Fig. 2). To assess whether a SNP-QTL significant in one analysis reached significance or suggestive thresholds in the other analyses, their *p*-values were compared. First, in the comparison between the lines-GWAS (Fig. 4a), most of the SNP-QTL detected via HRFI-GWAS had $-\log_{10}(p\text{-values})$ generally lower than the suggestive threshold of 3 in the LRFI-GWAS. Similar results were obtained comparing SNP-QTL of the LRFI-GWAS to their *p*-values with the HRFI-GWAS. For the SNP-QTL significant in the global-GWAS, the $-\log_{10}(p\text{-values})$ with the lines-GWAS were intermediate and exceeded the suggestive threshold in one of the lines for several QTL.

In addition, for the SNP-QTL corresponding to the QTL detected in the line analyses (HRFI-GWAS and LRFI-GWAS), the $-\log_{10}(p\text{-values})$ obtained in the

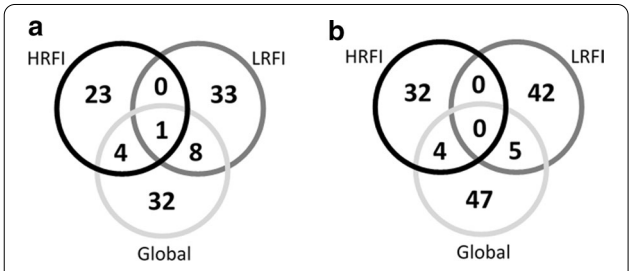

Fig. 3 Comparison of GWAS results obtained from Global-GWAS (Global), HRFI-GWAS (HRFI) and LRFI-GWAS (LRFI). **a** Comparison of the number of identical regions and **b** comparison of the number of identical QTL (trait x region)

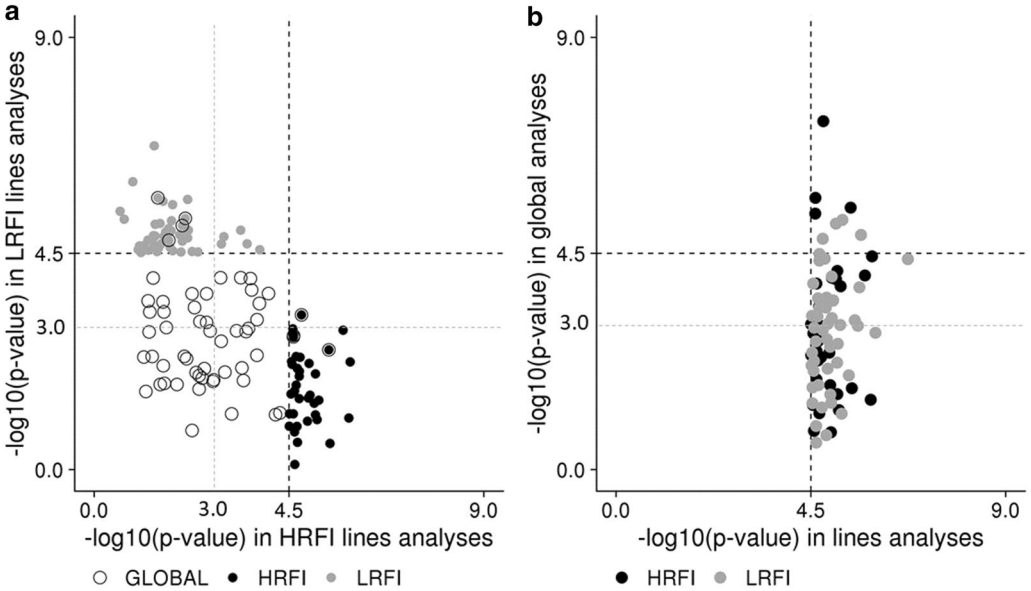

Fig. 4 Plot of the $-\log_{10}(p\text{-value})$ of the SNP-QTL. The $-\log_{10}(p\text{-value})$ are obtained in first case with the two lines analyses for all SNP-QTL detected for the lines or the global analyses (**a**), and in second case obtained with the global analysis for SNP-QTL detected with the GWAS performed per line (**b**)

global-GWAS were also low (Fig. 4b), with more than the half (56.6%) of the SNP-QTL having -$\log_{10}$(p-values) lower than 3.

Change in allele frequencies across generations

The allele frequencies of the SNP-QTL detected either in the global-GWAS or lines-GWAS were evaluated in G1 to G9 to reflect the informativeness of these GWAS (called G9 hereafter) and in G1. When the SNP-QTL was detected in the global-GWAS, all response animals were used to compute the frequencies; for SNP-QTL from the lines-GWAS, only the animals of the significant analysis (HRFI-GWAS or LRFI-GWAS) were used. The resulting frequency histograms are shown in Fig. 5. In G1 only, the distribution of the allelic frequencies of the SNP-QTL of the global-GWAS and that of the SNP-QTL of the lines-GWAS did not differ significantly (Fig. 5a). In G9, the distribution of the SNP-QTL allelic frequencies differed largely between the two types of analyses (Fig. 5b): 85.5% of the SNP-QTL of the global-GWAS remained in the same range of frequencies, between 0.2 and 0.6, whereas only 57.7% of the SNP-QTL of the lines-GWAS had allele frequencies within that range of values ($P < 0.001$ for a Chi^2 with 1 df, when comparing between the two types of analyses the number of SNP-QTL with frequencies between 0.2 and 0.6 with the number of SNP-QTL with other frequencies). In addition, 9% of the SNP-QTL of the lines-GWAS had a frequency higher than 0.6, whereas no marker reached such frequencies among the SNP-QTL of the global-GWAS.

In addition to the estimation of the global allelic frequencies, we evaluated if in each line the detected SNP-QTL in each type of analysis evolved differently. First, the differences in allele frequency between the HRFI and LRFI lines were estimated in the G1 generation (at the beginning of the selection) (Fig. 6). Regardless of the analysis (global- or lines-GWAS) in which the SNP-QTL was detected, initially more than 63% of the SNP-QTL showed small differences in line frequency (< 0.1) and less than 11% of the SNP-QTL showed a difference in line frequency greater than 0.2. These SNP-QTL were not particularly detected in one or the other type of analysis. Next, to better describe the changes in allele frequency across generations, frequencies of SNP-QTL from the global-GWAS and lines-GWAS were successively estimated in each line by adding data from the next generation to the previous generations: G1 allele frequencies were obtained from G1 individuals only, G2 allele frequencies were obtained from G1 and G2 individuals etc. Using the nine resulting frequencies computed for each line, a linear regression of the generation number on the allele frequencies was applied within line (Fig. 7). The comparison between lines of the regression coefficients of the allelic frequencies highlighted four distinct cases (Fig. 8). Altogether, the allelic frequencies of 4.5% of the SNP-QTL did not change with selection (Fig. 7a), 24.8% of the markers were co-selected in the two lines (Fig. 7b), 41.3% evolved in opposite directions in the two lines (divergence) (Fig. 7c), and 29.3% of the markers had frequencies that changed in one line only (17.3% in LRFI and 12% in HRFI) (Fig. 7d). Again no difference in the distribution of the SNP-QTL by category was identified in either type of analysis (p-value = 0.51 for a Chi^2 with 3 df).

For RFI in the two lines, four of the five detected QTL corresponded to regions that were selected in opposite directions in the lines, with strong differences in line frequencies: two RFI SNP-QTL showed differences in allelic frequency between lines greater than 0.2 in G1 and the other two RFI SNP-QTL showed large changes in allelic frequency (regression slope > 0.024/generation). To summarize the changes in SNP-QTL allele frequencies for each trait, an average evolution score between G1 and G9 was computed using the estimated evolution scores

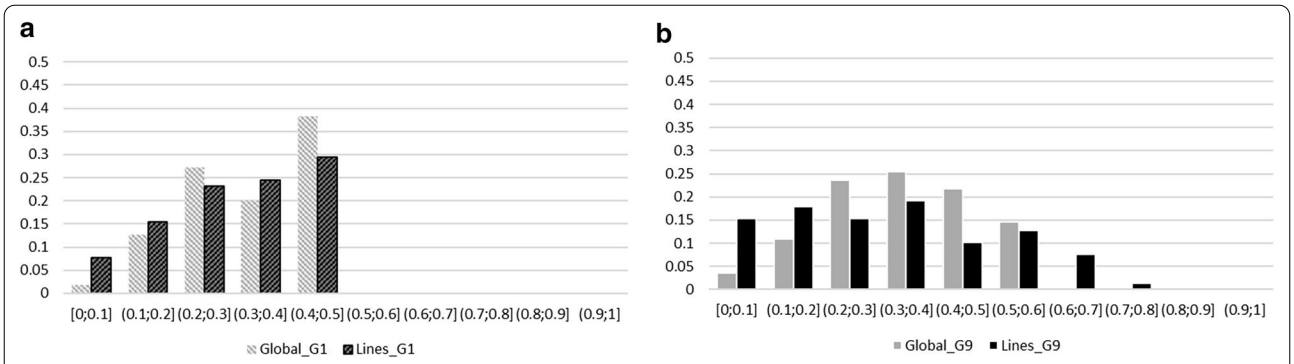

Fig. 5 Distribution of SNP-QTL allele frequencies of Global-GWAS (in grey) and Lines-GWAS (in black). Distribution representing individuals from the line of the significant analysis **a** in G1 generation (G1 individuals only) and **b** in G9 generation (G1 to G9 individuals)

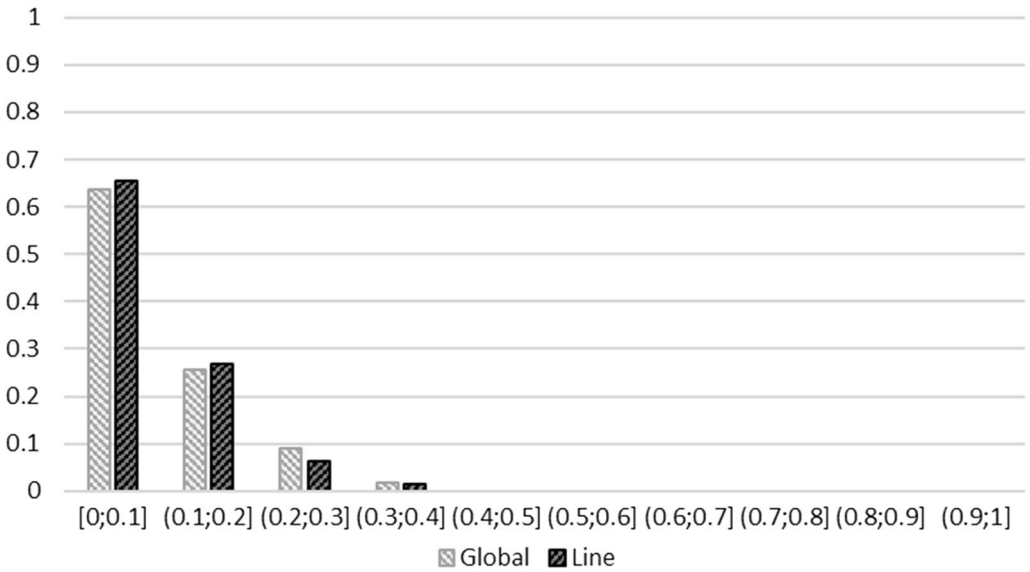

Fig. 6 Distribution of differences in allele frequencies between the lines. The differences in allele frequencies are the absolute values between lines for SNP-QTL resulting from the Global-GWAS and Lines-GWAS in G1

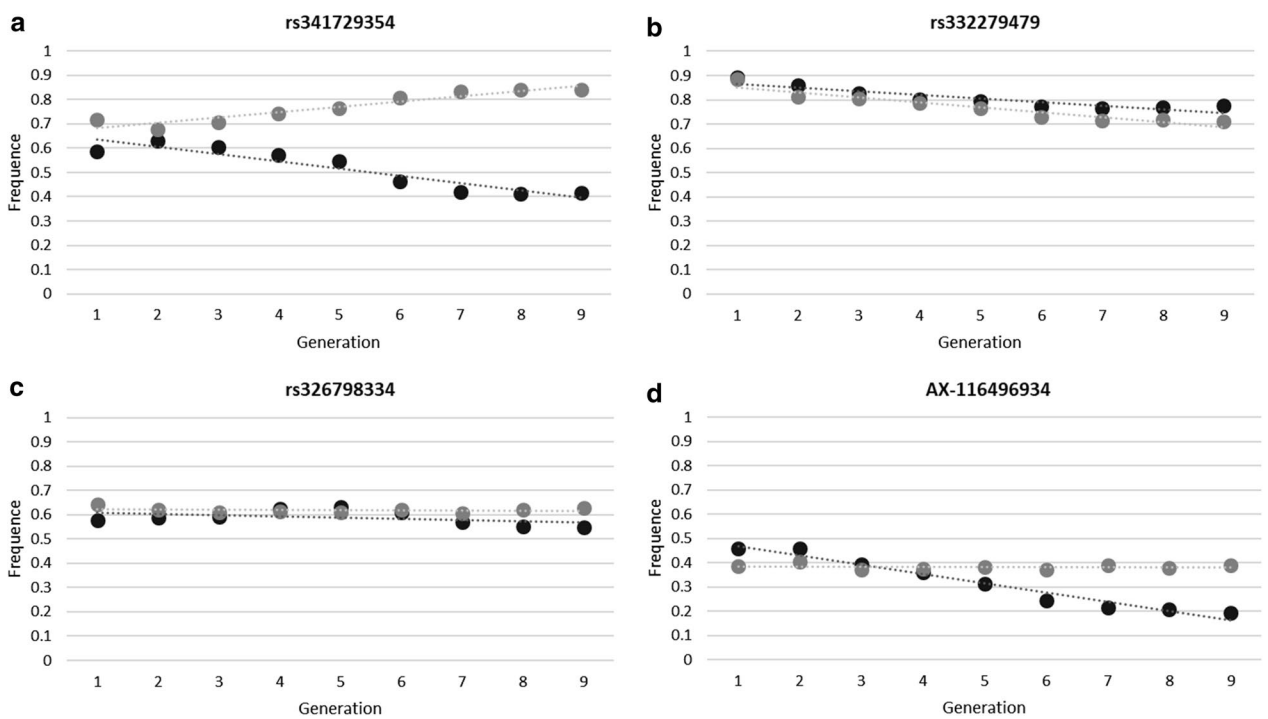

Fig. 7 Linear regression of the generation number on the allele frequencies computed in each line. Allele frequencies were estimated in the two lines by combining, for each generation, individuals of the generation n with the previous ones (animals from generation G1 to G n-1). Allelic frequencies evolutions are reported for SNP-QTL corresponding to **a** no-evolution, **b** co-evolution in both lines, **c** opposite-evolution, and **d** evolution only in one line, situations

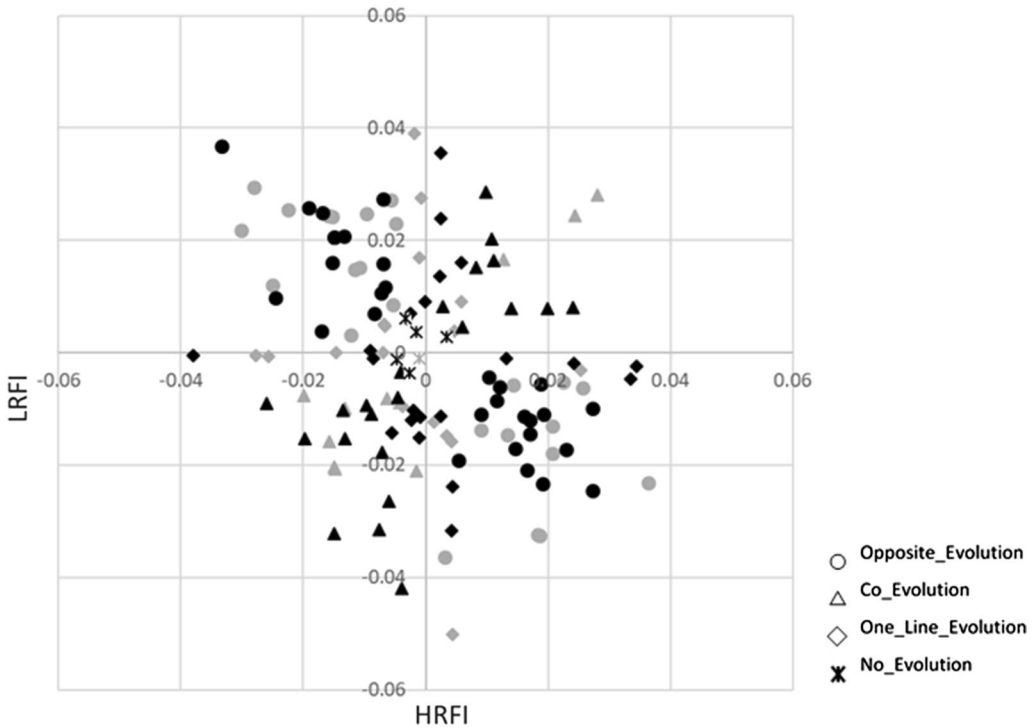

Fig. 8 Slopes of the linear regression equations of the allele frequencies on the nine generations. Slopes were calculated in each line, for all SNP-QTL identified with Global-GWAS (in grey) and Lines-GWAS (in black). Four situations (differentiated by different labels) were identified according to the significance of the slope (different from zero with $p < 0.05$ with a Wald test) in one or the two lines

of the different SNP-QTL detected for each trait. These averages were between 0.09 (Shoulder_W) and 0.35 (RFI). A correlation coefficient of 0.63 was then estimated between the genetic line differences in G9 computed previously for the 24 different traits [27] (Table 1) and these averages (Fig. 9).

Discussion

The objective of this study was to identify QTL that affect RFI and production traits in pig lines that have been divergently selected for RFI and to understand if the traits had different genetic backgrounds between the lines. By optimizing the genotyping to reach a sufficient power of detection of QTL in the full design and in the two lines, separately, QTL were detected for all traits and hypotheses about the trait genetic background in the two lines can be formulated.

Using average parental genotypes to detect QTL

While the use of SNP chips now enables the genotyping of an individual at a reasonable cost, the genotyping of a design comprising several thousands of individuals still represents a significant investment. In each generation of our design, at least two parities were produced, one to select future breeders, and one to control the responses to the selection on feed consumption, growth and meat quality traits via measurements at the slaughterhouse. After nine generations of selection, around 2500 "response animals" had phenotypes. These individuals have the advantage of having individual records for unmeasured traits in breeders (post-mortem measurements). To optimize the costs, we genotyped all 1632 breeders on MD SNP chips to exhaustively survey the segregating alleles in the design. In addition, the 32 main contributors to the design were chosen from the G0 sires and dams to be genotyped using the HD SNP chip, and an imputation step was carried out to have HD genotypes for all breeding individuals. The strong pedigree relationships in the design enabled a very good quality of HD imputation, since they help to better detect long haplotypes used to infer missing SNPs [28]. A second step was carried out, so that each response non-genotyped animal could have a genotype. Such imputation of non-genotyped animals has been used in cattle [29] as part of genomic evaluations to increase the size of reference populations. In cattle, the most common situation is to determine by imputation the genotypes of the dams of the bulls, knowing the genotypes of the maternal grandsire, one (or more) offspring and the sires with which they were mated [30]. In such cases, the strategy takes

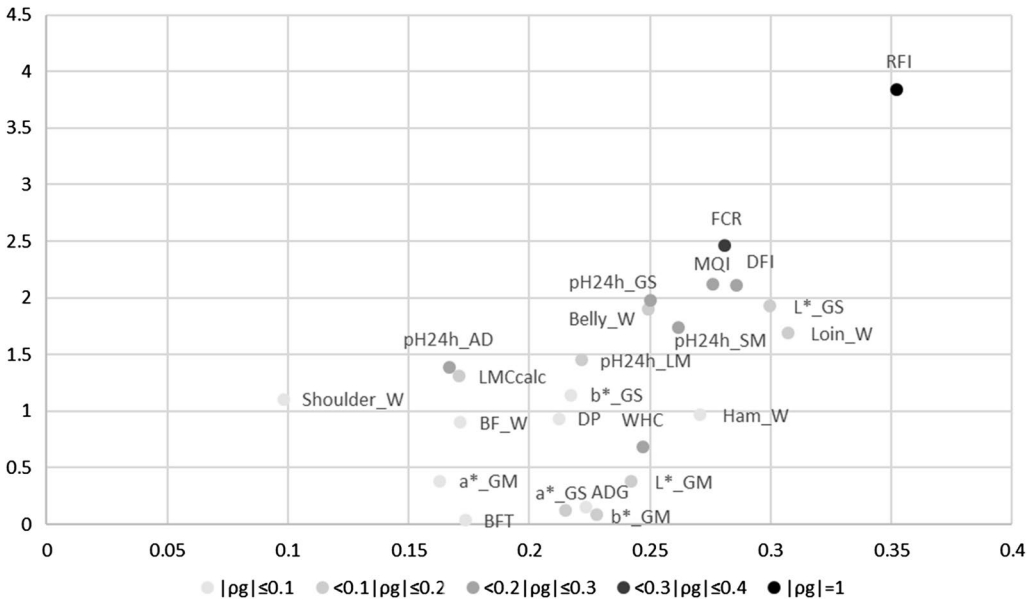

Fig. 9 Genetic differences in G9 between the two lines. The genetic differences were expressed in genetic standard deviations of the trait (σ_g) as a function of the average evolution of allelic frequencies in the QTL regions of the trait between the two lines. The magnitude of the genetic correlation between each trait and RFI is indicated with a grey gradient; DFI: daily feed intake; ADG: average daily gain; FCR: feed conversion ratio; RFI: residual feed intake; carcBFT: backfat thickness measured on carcass; a*_GM: a* measured on the *gluteus medius* muscle; a*_GS: a* measured on the *gluteus superficialis* muscle; b*_GM: b* measured on the *gluteus medius* muscle; b*_GS: b* measured on the *gluteus superficialis* muscle; L*_GM: L* measured on the *gluteus medius* muscle; L*_GS: L* measured on the *gluteus superficialis* muscle; pH24h_AD: pH 24 h after slaughter measured on the adductor femoris muscle; pH24h_GS: pH 24 h after slaughter measured on the *gluteus superficialis* muscle; pH24h_LM: pH 24 h after slaughter measured on the *longissimus dorsi* muscle; pH24h_SM: pH 24 h after slaughter measured on the *semimembranosus* muscle; WHC: water holding capacity of the *gluteus superficialis* muscle; MQI: meat quality index; LMCcalc: lean meat content of the carcass; DP: carcass dressing percentage; Belly_W: belly weight; BF_W: backfat weight; Ham_W: ham weight; Loin_W: loin weight; Shoulder_W: shoulder weight

advantage of family information (Mendelian rule of allele transmission) and combines it with allele frequencies and LD between markers at the population level. In our case, at each generation n, all response animals had both parents genotyped at generation n-1. Given these trio structures, an expected genotype at each position could be deduced from the genotypes of the parents using simple segregation rules: since the genotypes were coded as an allelic dosage for one reference allele, the genotype expectation for each offspring was simply the average of the genotypes of its two parents. As a result, 2426 animals with genotypes (predicted) and phenotypes were available for subsequent GWAS analyses.

Understanding the differences in the regions detected between analyses

The regions detected with each type of analysis (global- or lines-GWAS) were very different and only 10 among the 129 detected QTL were shared between global-GWAS and lines-GWAS. The SNP-QTL detected with the global-GWAS were far from reaching the threshold of significance in the lines-GWAS. Similarly, most of the SNP-QTL detected with the lines-GWAS were far from reaching the threshold of significance in the global-GWAS. Although the number of individuals included in the global-GWAS was more than twice that in the line analyses, the addition of individuals belonging to the other line seems to have reduced the power of detection of QTL segregating in the first line. Even if the allelic frequencies of the SNP-QTL detected in the global-GWAS or lines-GWAS were comparable in G1, they largely differed after nine generations of selection, i.e. more SNPs with low allele frequencies were identified with the lines-GWAS. The pedigree kinship matrix was used in the GWAS model to correct for the strong genomic structure of the population. Although this classical approach is successful to control type-I errors of the analyses, it also limits the power of detection of QTL in highly differentiated regions between lines, since their link with trait variability would be absorbed into the additive genetic component of the model. Thus, global-GWAS essentially allow the detection of regions that segregate at intermediate frequencies in both lines. As an alternative, the analyses carried out by line allow the detection of regions that are close to fixation with selection in one of the lines. From these results, it seems that the power of detection

related to allele frequencies in each line is the main difference between QTL-SNPs detected with the lines-GWAS and global-GWAS. Thus, given the power of the design, it is likely that the biological pathways involved in RFI variability in the two lines are similar, but with different contributions to the trait in each line, contrary to some previous hypotheses [10, 27].

Comparison with published regions

Among the five QTL detected for RFI, three regions were detected close to previously published RFI QTL. The region on SSC14 at 130–131 Mb is close to the region described by Do et al. [31] who proposed *G-protein-coupled receptor kinase 5* (*GRK5*) (129,114,449–129,343,412 bp) as a candidate gene. Wang et al. [32] reported that a GRK5 deficiency led to insulin resistance and hepatic steatosis, and to decreases in diet-induced obesity and adipogenesis in mice. At the position 131,181,710–131,579,703 bp, *FGFR2* (*fibroblast growth factor receptor 2*) could also be an interesting candidate gene. All four FGF receptors and several FGF ligands are present in the intestine and are key players in controlling cell proliferation, differentiation, epithelial cell restitution, and stem cell maintenance. *FGFR2* is expressed in the human ileum and throughout adult mouse intestine [33]. The second region closest to published RFI QTL is the 184–186 Mb interval on SSC13 near the QTL reported by Bai et al. [34] and Do et al. [31]. In this region, *TMPRSS15* (*transmembrane serine protease 15*) is an interesting candidate gene. This gene encodes an intestinal enzyme that is responsible for initiating the activation of pancreatic proteolytic proenzymes. It catalyzes the conversion of trypsinogen to trypsin, which in turn activates other proenzymes including chymotrypsinogen procarboxypeptidases and proelastases. *TMPRSS15* has been associated to enterokinase deficiency, a life-threatening intestinal malabsorption disorder characterized by diarrhea and failure to thrive [35]. On SSC17, two RFI QTL have been published by Do et al. [31] close to the *SOGA1* gene (*suppressor of glucose, autophagy-associated protein 1*, 40,020,107–40,098,992 bp) and by Onteru et al. [10] close to the *DOK5* gene (*docking protein 5*, 55,391,074–55,541,561 bp). These two QTL surround the region that we detected and could correspond to a unique QTL. At position 48,090,077–48,100,816 bp and at position 48,132,911–48,149,732 bp, respectively, *PLTP* and *ZNF335* are additional candidate genes. In humans, Coleman et al. [36] identified the region encoding ZNF335 as a susceptibility locus for the coeliac disease, a chronic immune-mediated disease triggered by the ingestion of gluten [36]. The PLTP (phospholipid transfer protein) transfers phospholipids from triglyceride-rich lipoproteins to high-density lipoprotein (HDL). In addition to regulating the size of HDL particles, this protein may be involved in the metabolism of cholesterol. *PLTP*-KO mice absorb less cholesterol than wild-type mice, and also have a deficient secretion in the intestine [37].

Potential pleiotropic effects

The large number of traits recorded in our design and the known genetic correlations between these traits [27] enable the detection of pleiotropic regions, i.e. regions that affect multiple traits. Among the four regions detected for FCR, no QTL co-localized with a RFI QTL. For the other traits, only two QTL were detected within 10 Mb of the RFI QTL: one QTL at 8 Mb influencing pH24h_LM on SSC9 between 1 and 2 Mb, and one QTL on pH24h_AD at 1 Mb of the QTL for RFI located at 113–114 Mb on SSC14. Compared to the previously published QTL regions for RFI, we identified only one QTL for DFI in the region described by Guo et al. [38] on SSC3 between 126 and 128 Mb. In spite of the reported correlations between these traits and RFI, among the 36 QTL detected in our study for DFI, MQI, WHC, pH24h_AD, pH24h_GS, and pH24h_SM, only one QTL co-located with the RFI QTL identified in our study or in previously published studies.

Changes in QTL allele frequencies and trait responses to selection

The allele frequencies of the majority of the detected regions changed between generations G1 and G9, with more than 70% of the regions for which SNP-QTL evolved in opposite directions or in one line only. However, the magnitude of the changes in allelic frequencies of the QTL regions varied among the traits, and was strongly correlated with previously reported line differences in G9 [27]. Indeed, the regions with the largest changes in allelic frequencies were detected for RFI, which was the trait used for selection. For the other traits, the higher the genetic correlation with RFI, the higher the variation in allelic frequency of the associated QTL regions. As a result, QTL that affect FCR, DFI and MQI showed the largest changes in allelic frequency with generations. The responses of QTL that affect meat quality traits are consistent with the high and early responses to selection previously detected in this experimental population for these traits [5]. Altogether, our analyses underline a clear relationship between the quantitative responses to selection of the traits and the changes in allelic frequencies in some QTL regions, which potentially point out to chromosomal regions that were selected during the experiment. Nevertheless, it is important to note that changes in allelic frequencies can also result from genetic drift. In such populations with a small effective size and strong directional selection, the

power of detection of signatures of selection using standard methodologies [39] can be low due to the major effect of genetic drift on the changes in allele frequencies. However, recently-developed new methods, based on genetic time series could provide new insights for the detection of regions under selection in small populations [40].

Conclusions

In this study, our aim was to characterize the molecular architecture of RFI in two lines that have been divergently selected for this trait. In addition to efficiently detecting known and new QTL regions, the combination of GWAS performed per line or simultaneously using all individuals allowed the identification of candidate regions on the genome and to understand how the genomes of both lines have evolved. Analyzing the allelic frequencies from G1 to G9, we identified that most of the differences in the results of QTL detection between the global or the two lines-GWAS were due to differences in informativity of the SNP-QTL in the two lines after nine generations of selection. Even if we cannot distinguish whether these evolutions in allelic frequencies are a direct effect of the directional selection or are due to drift, the regions detected can explain the responses to selection of different traits reported before. In addition, we conclude that the majority of the QTL regions followed divergent patterns in the lines, and that the same metabolic pathways were certainly involved in both lines. We identified several new regions that underlie RFI variability and propose new candidate genes that complement the data acquired in previously published analyses.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12711-021-00642-1.

Additional file 1: Table S1. Number of animals used for the analyses after quality control. Details of the number of animals before and after application of a filter on the call rate (CR) were given for chips (60K, 70K and 650K SNPs chips), imputation levels (MD/HD imputation) and average genotypes calculated from the genotypes of both parents (HD predicted).

Additional file 2: Table S2. Number of SNPs used for the analyses after quality control. Details of the number of SNPs before and after application of filters on the call frequency (CF) and the frequency of minor allele (MAF) were given for chips (60K, 70K and 650K SNPs chips), imputation levels (MD imputation and HD imputation) and average genotypes calculated from the genotypes of both parents (HD predicted).

Additional file 3: Figure S1. Correlations between true and imputed genotypes for animals genotyped on 60K, 70K or 650K SNPs chip. For each analysis, correlations were estimated setting 5000 SNPs as missing (5 batches of 1000 SNPs) on one chip among SNPs in common between the two arrays used. Animals are sorted and colored by generation. Correlations between true and imputed genotypes (a) for the 286 animals genotyped with the 60K SNPs chip using animals with 70K genotypes as reference population, and (b) for the 1346 animals genotyped with the 70K SNPs chip using animals with 60K genotypes as reference. (c) Correlations between true and imputed genotypes after imputation to 650K SNPs from the imputed medium density genotypes. (d) Correlations between true and imputed genotypes based on the leave-one-out cross-validation.

Additional file 4: Figure S2. Proportion of certain expected genotypes per animal, per SNP and in relation to the MAF of the SNPs. The proportion of certain genotypes corresponds to expected genotypes from parents which are homozygous for the same allele or homozygous for opposite alleles, and half of the genotypes from matings of two heterozygous parents were also taken into account. This proportion was studied per individual for the 66,988 SNPs of the 60K SNPs chip (a), per SNP for the 2426 pigs (b) and finally per SNP while taking into account the MAF of each SNP (c).

Additional file 5: Table S3. QTL regions detected with the three groups of association studies. These QTL regions were found from the full population (Global-GWAS) and from each line separately (HRFI-GWAS and LRFI-GWAS). DFI: daily feed intake; ADG: average daily gain; FCR: feed conversion ratio; RFI: residual feed intake; carcBFT: backfat thickness measured on carcass; a*_GM: a* measured on the *gluteus medius* muscle; a*_GS: a* measured on the *gluteus superficialis* muscle; b*_GM: b* measured on the *gluteus medius* muscle; b*_GS: b* measured on the *gluteus superficialis* muscle; L*_GM: L* measured on the *gluteus medius* muscle; L*_GS: L* measured on the *gluteus superficialis* muscle; pH24h_AD: pH 24 h after slaughter measured on the adductor femoris muscle; pH24h_GS: pH 24 h after slaughter measured on the *gluteus superficialis* muscle; pH24h_LM: pH 24 h after slaughter measured on the *longissimus dorsi* muscle; pH24h_SM: pH 24 h after slaughter measured on the *semimembranosus* muscle; WHC: water holding capacity of the *gluteus superficialis* muscle; MQI: meat quality index; LMCcalc: lean meat content of the carcass; DP: carcass dressing percentage; Belly_W: belly weight; BF_W: backfat weight; Ham_W: ham weight; Loin_W: loin weight; Shoulder_W: shoulder weight

Additional file 6: Figure S3. Manhattan plots for GWAS of RFI trait in global, HRFI line or LRFI line populations. The plot shows the −log10(p-values) for all SNPs in the analysis against their genomic position. Changes in color represent different chromosomes. The dashed line represents the threshold for genome wide significance (threshold of -log10(*p*-value) = 4.5).

Acknowledgements

The authors would like to thank (i) the experimental farm staff for data collection, samples management and breeding of the animals and (ii) both technology platforms, CRCT and Gentyane, for the genotyping.

Authors' contributions

ED performed the statistical analyses and wrote the first draft of the paper. YB and KF organized the data acquisition. ED and YL performed the imputation and quality control of the genotypic data. ED, AA, YL, HG and JR participated in the design of the study. JR and HG provided scientific supervision. All authors read and approved the final manuscript.

Funding

This study and the two first authors were financially supported by the French National Research Agency via the PIG_FEED and MicroFeed projects, under grants ANR-08-GENM-038 and ANR-16-CE20-0003.

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

All pigs were reared in compliance with national regulations and according to procedures approved by the French Veterinary Services at INRA experimental facilities. The care and use of pigs were performed following the guidelines edited by the French Ministries of High Education, Research and Innovation, and of Agriculture and Food (http://ethique.ipbs.fr/sdv/charteexpeanimale.pdf).

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]GenPhySE, Université de Toulouse, INRAE, ENVT, 31320 Castanet-Tolosan, France. [2]GenESI, INRAE, 17700 Surgères, France.

Received: 27 October 2020 Accepted: 28 May 2021
Published online: 14 June 2021

References

1. McGlone J, Pond W. Pig production: biological principles and applications. Florence: Thomson/Delmar Learning. 2003.
2. Soleimani T, Gilbert H. Evaluating environmental impacts of selection for residual feed intake in pigs. Animal. 2020;14:2598–608.
3. Webb AJ, King JWB. Selection for improved food conversion ratio on ad libitum group feeding in pigs. Anim Sci. 1983;37:375–85.
4. Koch RM, Swiger LA, Chambers D, Gregory KE. Efficiency of feed use in beef cattle. J Anim Sci. 1963;22:486–94.
5. Gilbert H, Bidanel J-P, Gruand J, Caritez J-C, Billon Y, Guillouet P, et al. Genetic parameters for residual feed intake in growing pigs, with emphasis on genetic relationships with carcass and meat quality traits. J Anim Sci. 2007;85:3182–8.
6. Cai W, Casey DS, Dekkers JCM. Selection response and genetic parameters for residual feed intake in Yorkshire swine. J Anim Sci. 2008;86:287–98.
7. Drouilhet L, Achard CS, Zemb O, Molette C, Gidenne T, Larzul C, et al. Direct and correlated responses to selection in two lines of rabbits selected for feed efficiency under ad libitum and restricted feeding: I. Production traits and gut microbiota characteristics. J Anim Sci. 2016;94:38–48.
8. 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.
9. 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.
10. Onteru SK, Gorbach DM, Young JM, Garrick DJ, Dekkers JCM, Rothschild MF. Whole genome association studies of residual feed intake and related traits in the pig. PLoS One. 2013;8:e61756.
11. Ding R, Yang M, Wang X, Quan J, Zhuang Z, Zhou S, et al. Genetic architecture of feeding behavior and feed efficiency in a Duroc pig population. Front Genet. 2018;9:220.
12. 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 Res. 2019;47:D701–10.
13. Sosa-Madrid BS, Santacreu MA, Blasco A, Fontanesi L, Pena RN, Ibáñez-Escriche N. A genomewide association study in divergently selected lines in rabbits reveals novel genomic regions associated with litter size traits. J Anim Breed Genet. 2020;137:123–38.
14. Daumas G. Taux de muscle des pièces et appréciation de la composition corporelle des carcasses. In: Proceedings of the 50th Journées de la Recherche Porcine: 6–7 February 2008; Paris. 2008;40:61–8.
15. Charpentier J, Monin G, Ollivier L. Correlations between carcass characteristics and meat quality in Large White pigs. In: Proceedings of the 2nd International Symposium on Conditions and Meat Quality of Pigs: 22-24 March 1971; Zeist. 1971.
16. Tribout T, Caritez J-C, Gogué J, Gruand J, Bouffaud M, Billon Y, et al. Estimation, par utilisation de semence congelée, du progrès génétique réalisé en France entre 1977 et 1998 dans la race porcine Large White : résultats pour quelques caractères de production et de qualité des tissus gras et maigres. In: Proceedings of the 34th Journées de la Recherche Porcine:3-5 February 2004; Paris. 2004;36:275–82.
17. Noblet J, Karege C, Dubois S, van Milgen J. Metabolic utilization of energy and maintenance requirements in growing pigs: effects of sex and genotype. J Anim Sci. 1999;77:1208–16.
18. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.
19. Warr A, Affara N, Aken B, Beiki H, Bickhart DM, Billis K, et al. An improved pig reference genome sequence to enable pig genetics and genomics research. GigaScience. 2020;9:1–14.
20. Sargolzaei M, Chesnais JP, Schenkel FS. A new approach for efficient genotype imputation using information from relatives. BMC Genomics. 2014;15:478.
21. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012;44:821–4.
22. Aliakbari A, Delpuech E, Labrune Y, Riquet J, Gilbert H. The impact of training on data from genetically-related lines on the accuracy of genomic predictions for feed efficiency traits in pigs. Genet Sel Evol. 2020;52:57.
23. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genet. 2013;9:e1003264.
24. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999;55:997–1004.
25. Gao X. Multiple testing corrections for imputed SNPs. Genet Epidemiol. 2011;35:154–8.
26. Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: a review. Plant Methods. 2013;9:29.
27. Gilbert H, Billon Y, Brossard L, Faure J, Gatellier P, Gondret F, et al. Review: divergent selection for residual feed intake in the growing pig. Animal. 2017;11:1427–39.
28. Ullah E, Mall R, Abbas MM, Kunji K, Nato AQ, Bensmail H, et al. Comparison and assessment of family- and population-based genotype imputation methods in large pedigrees. Genome Res. 2019;29:125–34.
29. Bouwman AC, Hickey JM, Calus MP, Veerkamp RF. Imputation of non-genotyped individuals based on genotyped relatives: assessing the imputation accuracy of a real case scenario in dairy cattle. Genet Sel Evol. 2014;46:6.
30. Pimentel EC, Wensch-Dorendorf M, König S, Swalve HH. Enlarging a training set for genomic selection by imputation of un-genotyped animals in populations of varying genetic architecture. Genet Sel Evol. 2013;45:12.
31. Do DN, Ostersen T, Strathe AB, Mark T, Jensen J, Kadarmideen HN. Genome-wide association and systems genetic analyses of residual feed intake, daily feed consumption, backfat and weight gain in pigs. BMC Genet. 2014;15:27.
32. Wang L, Shen M, Wang F, Ma L. GRK5 ablation contributes to insulin resistance. Biochem Biophys Res Commun. 2012;429:99–104.
33. Danopoulos S, Schlieve CR, Grikscheit TC, Alam DA. Fibroblast growth factors in the gastrointestinal tract: twists and turns. Dev Dyn. 2017;246:344–52.
34. Bai C, Pan Y, Wang D, Cai F, Yan S, Zhao Z, et al. Genome-wide association analysis of residual feed intake in Junmu No. 1 White pigs. Anim Genet. 2017;48:686–90.
35. Holzinger A, Maier EM, Bück C, Mayerhofer PU, Kappler M, Haworth JC, et al. Mutations in the Proenteropeptidase gene are the molecular cause of congenital enteropeptidase deficiency. Am J Hum Genet. 2002;70:20–5.
36. Coleman C, Quinn EM, Ryan AW, Conroy J, Trimble V, Mahmud N, et al. Common polygenic variation in coeliac disease and confirmation of ZNF335 and NIFA as disease susceptibility loci. Eur J Hum Genet. 2016;24:291–7.
37. Liu R, Iqbal J, Yeang C, Wang DQ, Hussain MM, Jiang X-C. Phospholipid transfer protein–deficient mice absorb less cholesterol. Arterioscler Thromb Vasc Biol. 2007;27:2014–21.
38. Guo YM, Zhang ZY, Ma JW, Ai HS, Ren J, Huang LS. A genomewide association study of feed efficiency and feeding behaviors at two fattening stages in a White Duroc × Erhualian F2 population. J Anim Sci. 2015;93:1481–9.
39. 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.
40. Paris C, Servin B, Boitard S. Inference of selection from genetic time series using various parametric approximations to the Wright-Fisher model. G3 (Bethesda). 2019;9:4073–86.

Publisher's Note