Approche éco-anatomique du bois de vigne (Vitis vinifera L.) pour une meilleure connaissance de l'histoire de la viticulture en Méditerranée nord-occidentale

Katia Feve

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE

ECOLE PRATIQUE DES HAUTES ETUDES

Sciences de la Vie et de la Terre

MEMOIRE

Présenté par

Katia FEVE

Pour l'obtention du diplôme de l'Ecole Pratique des Hautes Etudes

CARTOGRAPHIE FINE ET CARACTERISATION D'UN QTL LOCALISE SUR LE CHROMOSOME 1 PORCIN INFLUENCANT L'ADIPOSITE DES ANIMAUX.

Soutenance proposée le 27 Novembre, devant le jury suivant :

Président :	Thierry Dupressoir (DE)
Rapporteur :	Alain Ducos (PR)
Examinateur :	Frédéric Lecerf (MC)
Tutrice scientifique :	Juliette Riquet (DR)
Tutrice pédagogique :	Christelle Lasbleiz (MC)

Mémoire préparé sous la direction de :

Juliette RIQUET
Equipe GenEpi (Génétique et Epigénétique des espèces utilisées en croisements).
Unité de Génétique, Physiologie et Systèmes d'Elevage UMR1388
Directeur : Xavier FERNANDEZ

Christelle LASBLEIZ
Equipe Vieillissement Cérébral et Maladies Neurodégénératives
Unité de Mécanismes Moléculaires dans les Démences Neurodégénératives UMR_S1198
Directeur : Dr Jean-Michel VERDIER

Table des matières

Remerciements

Voilà donc une expérience incroyable qui se termine et il m'est impossible de clôturer ces 4 années de diplôme sans vous dire à quel point je vous remercie… Si je tiens à remercier toutes les personnes qui m'ont permis d'accomplir cette aventure, ce n'est pas juste par tradition, mais bien pour tout ce qu'elles m'ont apporté durant ces années. Que ce soit d'un point de vue professionnel, scientifique ou personnel, tout cela n'aurait pas été possible sans l'aide, l'encadrement et le soutient de beaucoup de personnes !

Merci à tous les membres du jury qui ont accepté d'évaluer ce travail. Un grand merci à Alain Ducos et à Frederic Lecerf d'avoir accepté d'être rapporteur et examinateur de ce mémoire. Je remercie également Monsieur Thierry Dupressoir d'avoir accepté de présider la soutenance de ce mémoire.

Je veux aussi adresser mes vifs remerciements à Madame Christelle Lasbleiz, qui a accepté d'être ma tutrice pédagogique malgré un sujet sur les QTL, ainsi que pour ces nombreux conseils lors de la relecture du CCR ou de ce mémoire.

A mon équipe :

Tout d'abord un grand Merci à Juliette, nous avons démarré ensemble et depuis les premiers jours tu m'as toujours poussé à aller au-delà de mes limites.
Merci de m'avoir confié les rênes du QTL du chromosome 1 et d'avoir cru en moi, là ou moi-même je n'y croyais pas. Merci de ta patience et de tes conseils lors de la rédaction de ce mémoire. J'ai eu des moments de doute car ce n'est pas évident de se sentir à la hauteur face à toi. Cette expérience aura été plus qu'enrichissante dans beaucoup de domaines.
Sans toi, je ne serais certainement pas allée au bout de cette aventure. Je te souhaite à mon tour bonne chance dans ce nouveau défi que tu as décidé de relever…Je ne doute pas que là encore, tu réussiras.

Je tiens également à remercier particulièrement Pitou, pour ton soutien amical. C'est grâce à toi que j'ai mis un pied dans ce labo et c'est également toi qui m'a permis de ne plus le quitter. Travailler à tes côtés est une véritable chance et source d'inspiration et cela depuis le premier jour. Je te remercie également pour ta disponibilité et tes nombreuses relectures de ce mémoire.

Merci aux bio-informaticiens, tout spécialement Patrice et Philippe, qui ont eu la patience et la gentillesse de toujours répondre à mes questions, de m'expliquer chaque étape qui m'ont permis d'analyser l'ensemble de ces données. Et un grand merci à Julien, qui est à l'origine de la production de ces données de séquences, notamment des nombreuses librairies qu'il a été amené à faire au cours de ce projet.

Enfin, merci à toutes les personnes de mon équipe et de mon laboratoire, qui ont rendu cette expérience possible, et à tous ceux qui m'ont aidé de près ou de loin dans cette aventure !

A ma famille :

A mes parents. Papa merci de m'avoir transmis cette volonté de ne jamais rien lâcher, de m'avoir laissé grandir dans un cadre où ma curiosité a pu s'épanouir sans limite. A ma maman, pour son amour, sa patience et son soutien indéfectible… Je ne serais jamais assez reconnaissante pour tout ce que tu as fait, et continue de faire pour moi. Ce mémoire aura été aussi l'occasion de me rendre compte que nous partageons finalement une passion commune. En effet, la génétique et l'étude des langues anciennes ont un même objectif, nous recherchons toutes les 2 à déchiffrer et comprendre un alphabet particulier…

Enfin, mes pensées les plus tendres vont bien sûr à mon épouse et à mes 2 zozios, qui ont été ma force et mon énergie au cours de ces 4 années.

Enfin, je ne pourrais pas terminer ces remerciements sans une pensée particulière pour Mr Terzian. Il a suscité mon intérêt pour la bio-informatique, en comparant les alignements de séquence avec les Souhern-blot…Nos échanges m'ont permis de comprendre que la bio-informatique n'était pas un domaine si hermétique, mais c'était avant tout un outil puissant pour répondre à un très grand nombre de questions scientifiques. Cela nécessite la même rigueur que lors de la réalisation d'une expérience. Je pense que c'est en partie grâce à ses enseignements, que ce projet a pris cette orientation et comporte une aussi importante partie bio-informatique.

A Edwige, Quentin et Lounha

« Il faut toujours viser la lune, car même en cas d'échec, on atterrit dans les étoiles »

Oscar Wilde

Listes des illustrations

Listes des tableaux

Liste des Abréviations

ADN	Acide Désoxyribo-Nucléique
ARN	Acide Ribo-Nucléique
BAC	Bacterial Artificial Chromosome
BC	Back-Cross
BET	Bromure d'Ethidium
BLUP	Best Linear Unbiased Predictor
cM	centiMorgan
DL	Déséquilibre de Liaison
ELD	Epaisseur de Lard Dorsal
EST	Expressed Sequence Tag
FAO	Food and Agriculture Organization of the United Nations
FISH	Fluorescence In Situ Hybridization
GMQ	Gain Moyen Quotidien
GPR107	G Protein-coupled Receptor 107
HMCN2	Hemicentin 2
HSA	Homo sapiens
IBD	Identical By Descent
IFIP	Institut de la Filière Porcine
IGF2	Insulin-like Growth Factor
INRA	Institut National de la Recherche Agronomique
LOD	Logarithm of the odds
LR	Landrace Français
LW	Large White
LWF	Large White lignée Femelle
LWM	Large White lignée Mâle
Mb	Mégabase
MS	Meishan
NCS1	Neuronal Calcium Sensor 1
OSP	Organisme de la Sélection Porcine
PCR	Polymerase Chain Reaction
PI	Piétrain
PIC	Polymorphic Information Content
QTL	Quantitative Trait Locus
RN	Rendement Napole
SAM	Sélection Assistée par Marqueur
SNP	Single Nucleotide Polymorphism
SSC	Sus Scrofa
TA (SC)	Tissu Adipeux (Sous-Cutané)
TAE	Tris Acetate EDTA
TBE	Tris Borate EDTA
TVM	Taux de Viande Maigre
UTR	UnTranslated Region

INTRODUCTION BIBLIOGRAPHIQUE

CHAPITRE I : DONNEES BIBLIOGRAPHIQUES

1 LA FILIERE PORCINE

1.1 La production de viande porcine : une demande en constante augmentation

Les animaux domestiques représentent l'une des principales sources d'apports en protéines de l'alimentation humaine. La consommation mondiale de viande a augmenté de 65% en un demi-siècle. Elle est passée de 25 kg/hab./an en 1970 à 42 kg/hab./an aujourd'hui. Les données de la FAO (*Food and Agriculture Organization)* disponibles sur la production de viande permettent de suivre son évolution de 1961 à nos jours. La production mondiale de viande porcine est ainsi passée de 20 millions de tonnes dans les années 1960 à 115,5 millions de tonnes aujourd'hui ; la production de volaille, qui était à moins de 10 millions de tonnes, atteint désormais presque 110 millions de tonnes, la Chine en produisant une très grande partie ; la production de viande bovine est passée d'environ 30 millions à 68 millions de tonnes, le Brésil en étant le premier producteur et exportateur mondial. La production de viande dans le monde est estimée à 305 millions de tonnes dont 37,7 % de viande porcine, 35,5 % de volaille et 22,2 % de viande bovine (année 2014, source FAO). La production mondiale de viande a plus que quintuplé entre 1960 et 2000 et la croissance se poursuit depuis cette date à un rythme très soutenu, même si elle tend à stagner, voire à diminuer, dans les pays développés (Figure 2).

Figure 1 : Produits d'origine animale consommés dans le monde par an.

(https://www.planetoscope.com/elevage-viande/1235-consommation-mondiale-de-viande.html).

Globalement, les pays dont la consommation de viande de porc est la plus importante sont également ceux qui en produisent le plus. La Chine est le premier producteur du monde, avec plus de 50 millions de tonnes produites en 2010. Sa production a été multipliée par trente en cinquante ans.

Les deuxième et troisième plus grands producteurs de viande porcine sont l'Europe avec 22,25 millions et les Etats-Unis avec 10 millions de tonnes. Au sein de l'Europe, la France est le troisième producteur de porc, avec 2,3 millions de tonnes produites chaque année, derrière l'Allemagne et l'Espagne, qui produisent 4,9 et 3,5 tonnes respectivement par an. Depuis une dizaine d'année la production porcine française est stable alors qu'elle s'est développée fortement en Allemagne (+26%) et en Espagne (+17%). Le principal bassin de production en France est situé en Bretagne, et représente 58% de la production porcine française.

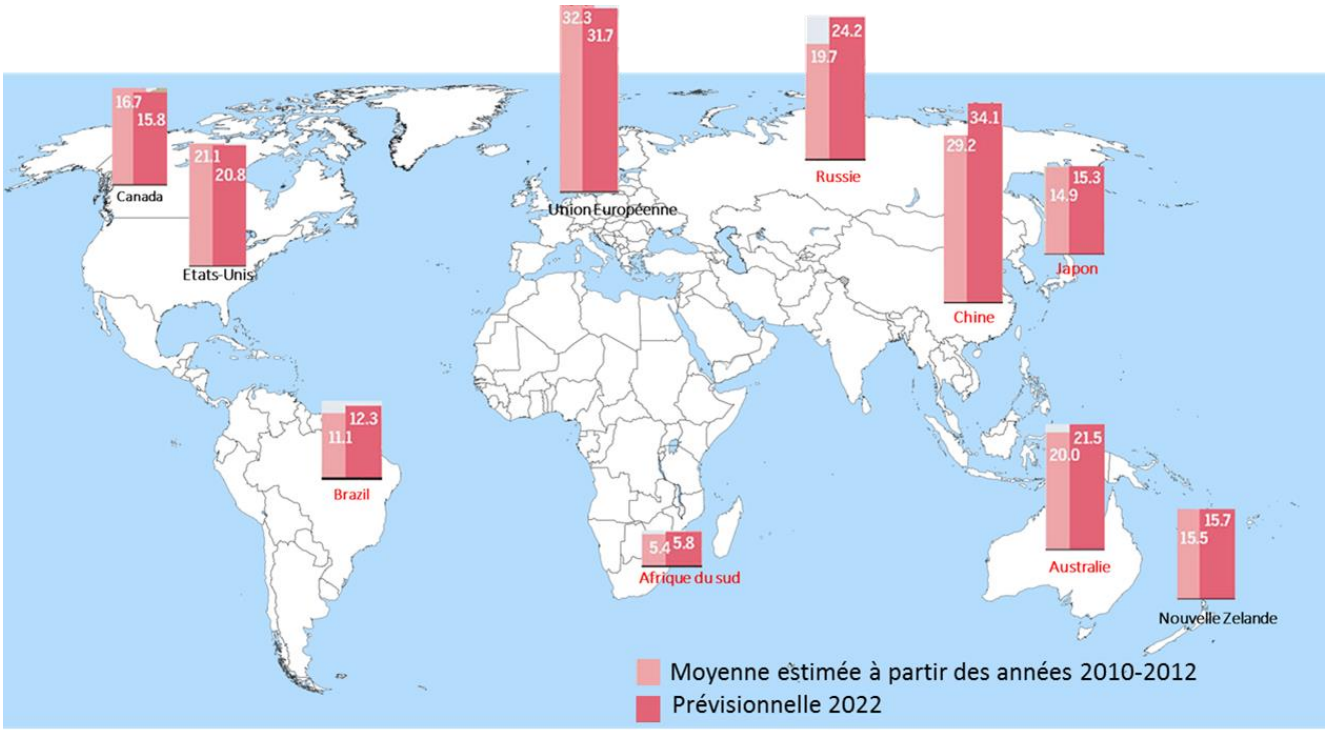


Figure 2 : Consommation mondiale de viande de Porc en kg par habitant.

D'après https://www.alimentarium.org/fr/magazine/soci%C3%A9t%C3%A9/une-folle-envie-de-viande.

Le développement économique de nombreux pays conduit à une augmentation de la consommation de viande par les populations et notamment de produits du porc (Figure 2). En effet, la FAO estime que la demande en viande devrait progresser de 200 millions de tonnes entre 2010 et 2050, soit pratiquement un doublement des volumes actuellement produits (Synthèse FranceAgrimer, 2011). Les consommations de viande de volailles et de porc devraient connaître les plus fortes croissances.

Pour pouvoir atteindre ces niveaux de production qui ne cessent d'augmenter à l'échelle mondiale, les différents maillons de la filière font l'objet d'évolution et d'amélioration constante. Parmi ces maillons, la sélection animale a permis depuis de très nombreuses années la production d'animaux répondant au mieux à la demande des producteurs, transformateurs et consommateurs.

1.2 La production et la sélection

La production porcine est organisée selon un schéma pyramidal à 3 niveaux (Figure 3). La sélection est uniquement réalisée au premier niveau de la pyramide et vise à créer le progrès génétique au sein de races pures, en sélectionnant les meilleurs animaux reproducteurs.

Le second étage consiste à une multiplication des animaux issus de la sélection, afin de constituer les populations parentales hybrides. Cette pratique du croisement permet de bénéficier d'une part de la complémentarité entre races, et d'autre part de l'effet d'hétérosis (supériorité des croisés par rapport à la moyenne des populations parentales) sur certains caractères

Enfin le dernier étage de la pyramide correspond à l'étape de production des porcs charcutiers commercialisés pour la production de viande.

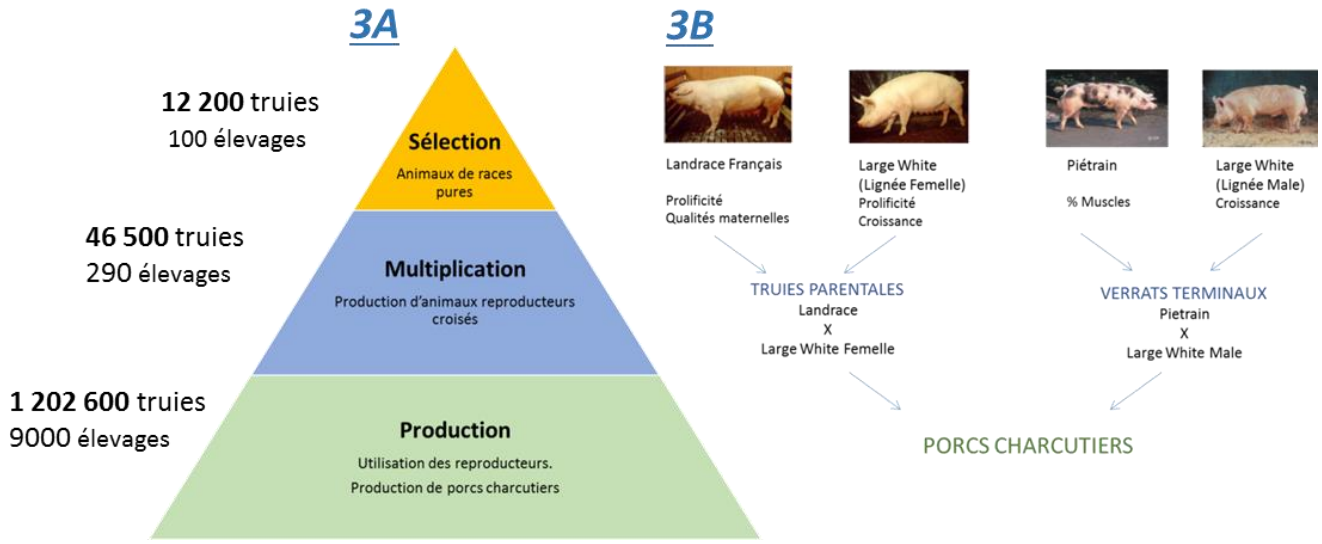


Figure 3 : Organisation pyramidale de la production porcine.

3A : Effectifs de truies et nombre d'élevages concernés par chaque étage de la production. 3B : Exemple possible de l'utilisation de 4 races pures dans un plan de croisement pour obtenir des porcs charcutiers ; NB : A l'heure actuelle le verrat terminal le plus utilisé est un verrat de race pure Piétrain.

La sélection vise à améliorer les performances des animaux en choisissant à chaque génération comme reproducteurs, les animaux les plus performants pour les caractères d'intérêt. Si pendant de très nombreuses années (voire des siècles) ce choix a été réalisé par l'observation des performances propres des individus, des modèles de génétique quantitative développés au cours du XXème siècle ont permis le développement d'outils plus précis et performants. Ces outils ont alors permis de déterminer la valeur génétique des animaux afin de classer les candidats à la sélection selon un critère unique appelé indice (ou index) à la sélection combinant les différents caractères pondérés pour leur importance relative.

Initialement, le calcul de cet indice ne prenait en compte que les performances propres de l'individu. Avec l'amélioration des méthodes d'évaluation génétique et notamment la généralisation de la méthode BLUP (Best Linear Unbiased Prediction, Henderson, 1959) il est maintenant possible d'obtenir avec une plus grande précision l'estimation de la valeur génétique d'un animal. En effet, ce modèle prend non seulement en compte les phénotypes de l'individu mais aussi toutes les performances des individus apparentés (ascendants, collatéraux, descendants) et les effets de l'environnement. Ces nouvelles méthodes ont permis un fort progrès génétique depuis les années 80. Il est cependant important de souligner que les caractères dont les progrès sont les plus importants sont les caractères (1) dont la contribution génétique par rapport à l'effet environnemental est forte (notion exprimée par la valeur d'héritabilité h2), (2) dont le poids dans l'index est important et (3) pour lesquels un phénotype non invasif et simple à mesurer permet de disposer d'une mesure pour un très grand nombre d'individus dans la population.

Afin d'optimiser ces programmes de sélection, les différentes races ou lignées porcines ont été sélectionnées de façon « spécialisée », à orientation de « type femelle », ou de « type mâle ». Dans les lignées femelles on recherchera généralement à concilier autant des caractères de reproduction que des caractères de production. En revanche pour les lignées mâles uniquement les caractères de production seront pris en compte (Institut technique du porc, France, 2013). Au fur et à mesure du temps, afin de répondre au mieux à la demande, les index sont réévalués afin d'intégrer de nouveaux caractères d'importance économique et de faire évoluer les pondérations : pour les caractères de reproduction, le « nombre de petits sevrés » a été ajouté au « nombre de nés vivants » ; pour les caractères de production, la consommation moyenne journalière (CMJ) a été ajoutée aux mesures de croissance. Le choix des races pour la production des parents des porcs charcutiers se fera donc en

fonction des objectifs de sélection et des caractéristiques de chacune des races pures grand-parentales (Figure 4).

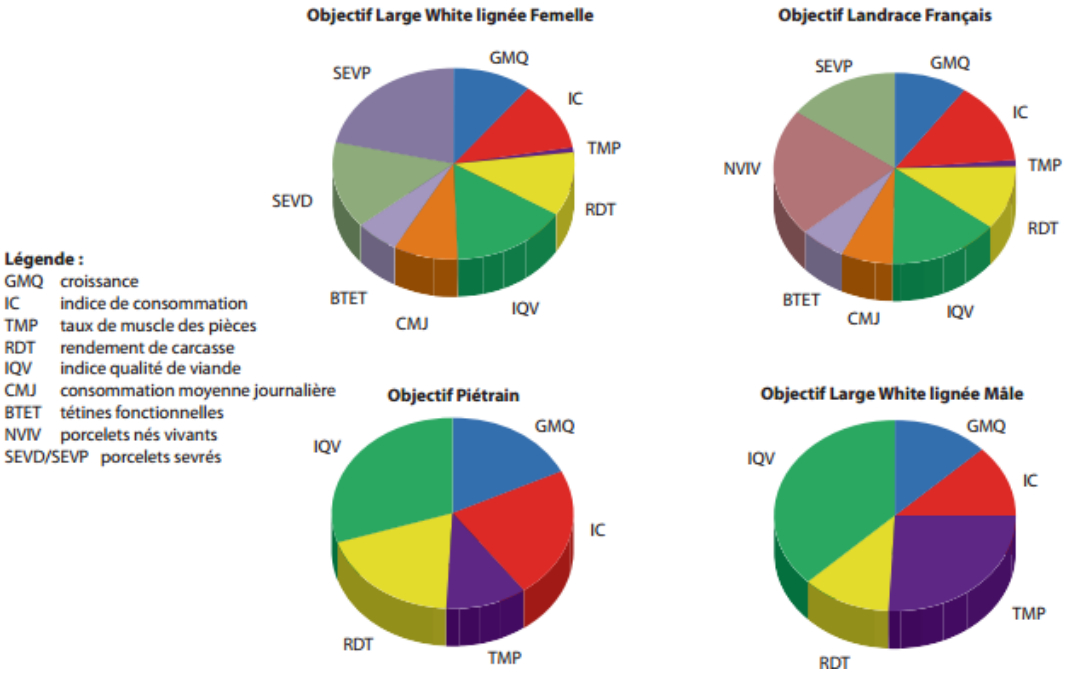


Figure 4 : Pondérations accordées aux différents objectifs de sélection pour les 4 races sélectionnées de manière collective en France.
Figure issue du Mémento de l'éleveur Porc, 7ieme édition, 2013.

Parmi les 350 races qui ont été répertoriées dans le monde (Mason, 1988), 4 grandes catégories peuvent être distinguées selon leurs performances de production et/ou de reproduction (Legault, 1978) :

- Les races mixtes présentant de bonnes aptitudes à la fois en production (croissance) et en reproduction. Le Large White et le Landrace d'origine européenne et la race américaine Duroc sont des exemples de races mixtes.
- Les races spécialisées en production (Piétrain) ; elles se caractérisent par un bon niveau de croissance et d'efficacité alimentaire mais surtout par une « musculature » exceptionnelle ; elles sont donc généralement utilisées en voie paternelle dans le croisement terminal.
- Les races très prolifiques, essentiellement des races chinoises (Jia Xing ou Meishan) ; elles se distinguent par une précocité sexuelle exceptionnelle (3 à 4 mois contre 6 à 8 mois) et une taille de portée pouvant atteindre 13 à 15 porcelets nés-vivants par portée, de bonnes aptitudes maternelles mais des performances de production très faibles.
- Les races locales (Gascon, Corse, Cul noir du limousin, …) ; elles sont également moins intéressantes pour des caractères de production ou de reproduction mais elles sont très bien adaptées à des milieux très difficiles. Elles sont en général utilisées pour des productions locales spécifiques haut de gamme.

En parallèle de cette sélection, certaines OSP (Organisme de Sélection Porcine) ont développé depuis les années 1970 leurs propres lignées synthétiques à partir de la fusion de 2 ou 3 races existantes, puis sélectionnées comme de nouvelles races (P76, Naima, Taizumu, Duochan…). Après plusieurs générations d'intercrosses, lorsqu'une certaine homogénéité est installée, on considère qu'une nouvelle lignée est créée. Ce mode de croisement est particulièrement intéressant quand on recherche pour plusieurs caractères une position intermédiaire entre les deux races parentales. Dans le cas des croisements entre races porcines chinoises et européennes, ce type de croisement a pour objectif de combiner les avantages de la lignée Meishan, c'est-à-dire

une très forte prolificité, tout en conservant les bonnes performances de production de la lignée Large White. Elle est actuellement utilisée dans les plans de croisement pour la voie maternelle (Milan *et al*., 2003).

1.3 L'apport de la génétique moléculaire à la filière porcine

1.3.1 Le développement de la génétique moléculaire

Si la généralisation du BLUP, à partir des années 80, a révolutionné la sélection animale, des évolutions technologiques dans le domaine de la biologie moléculaire dans les années 90 sont apparues comme sources d'une nouvelle révolution. Ces développements importants en biologie moléculaire sont (1) la mise au point de la technique d'amplification en chaine par la polymérase (PCR) et (2) la découverte d'un nombre considérable de polymorphismes dans l'ADN.

Alors que les modèles de génétique quantitative supposent un déterminisme polygénique et nécessitent la mesure de performances sur un grand nombre d'animaux (représentant une opération excessivement lourde et coûteuse), l'essor de la génétique moléculaire a permis d'envisager l'identification de l'ensemble des variations génétiques (mutations de l'ADN) sources de la variabilité des performances observées. Une nouvelle organisation de la sélection et de l'amélioration génétique devenait envisageable : la caractérisation de chaque animal à ces mutations allait permettre de connaitre sa valeur génétique sans aucune mesure phénotypique.

Cependant, contrairement à l'essor des travaux réalisés sur le génome humain, peu de données sur les génomes des espèces animales étaient disponibles au début des années 1990 ; compte tenu des difficultés techniques et du faible nombre d'équipes engagées, une dizaine de laboratoires Européens ont alors choisi de collaborer dans le cadre du projet PiGMaP-Bridge (Archibald *et al*., 1995). Cette collaboration avait pour principaux objectifs (1) de définir un premier réseau de marqueurs génétiques de type microsatellites espacés en moyenne de 20 cM afin d'établir une carte génétique à partir de familles de références (Gellin and Grosclaude, 1991), et (2) d'améliorer nos connaissances sur le contenu et l'organisation des gènes dans le génome.

1.3.2 Les domaines d'application

Au moins 4 domaines de la génétique animale ont pu bénéficier de ces avancées en génétique moléculaire et notamment du développement des marqueurs génétiques :

1.3.2.1 Caractérisation de la diversité génétique

Le développement et l'utilisation des marqueurs microsatellites, très polymorphes et en général localisés dans des régions neutres du génome (c'est-à-dire non soumis à la sélection) ont fait d'eux des outils de choix pour la caractérisation de la variabilité génétique entre races. En effet, une bonne connaissance génétique de ces populations permet de concevoir des méthodes de gestion et de conservation de la diversité plus efficaces. Une étude réalisée sur 11 races européennes en provenance de six pays européens, et incluant un petit échantillon de sangliers, a été menée en 1999 (Laval *et al*., 2000) pour estimer la biodiversité de l'espèce. Cette étude a démontré l'importance des races locales françaises (Basque, Gascon, Limousin et Bayeux). En effet elles contribuent à expliquer plus de la moitié de la diversité génétique totale. Cela indique la valeur potentielle de ces races locales, menacées de disparition, dans le maintien de la diversité génétique d'une espèce.

1.3.2.2 Contrôle de filiation

Ces marqueurs ont été aussi très utilisés dans le cadre de la vérification des filiations. L'évaluation de la valeur génétique d'un individu prend en compte l'ensemble des relations de parentés et des performances disponibles. La fiabilité de ces valeurs sera donc très fortement altérée si une partie des liens de parenté pris en compte est erronée. Actuellement un contrôle de filiation à l'aide d'un set de marqueurs moléculaires est réalisé

pour l'ensemble des candidats à la sélection. Ce contrôle est également nécessaire pour la certification des pedigrees avant de pouvoir intégrer une nouvelle lignée au sein des livres généalogiques.

1.3.2.3 La recherche de gènes majeurs et mise en évidence de QTL

Le développement d'outils d'analyse tout génome et les cartes génétiques ont permis la mise en place de très nombreux programmes de recherche des régions du génome (voire des gènes eux-mêmes) intervenant dans la variabilité des caractères quantitatifs (QTL), tels que la croissance, la composition de la carcasse (tissus musculaires et tissus adipeux), la qualité de la viande (paramètres technologiques et androstérone), les caractères de reproduction (fertilité, prolificité, taux d'ovulation…).
L'identification de ces régions à l'aide de marqueurs moléculaire a pu conduire à 2 nouveaux champs d'application pour l'amélioration génétique :

- D'une part de concevoir une sélection assistée par marqueur. Cette approche est décrite dans le paragraphe suivant.
- D'autre part, d'envisager la caractérisation moléculaire des gènes sous-jacents à la variabilité des caractères. Cette seconde approche sera détaillée dans la seconde partie bibliographique.

1.3.2.4 L'introgression et la sélection assistée par marqueurs (SAM)

Une application majeure envisagée avec l'aboutissement des travaux de génétique moléculaire est de pouvoir caractériser le génotype des animaux dès leur naissance grâce aux mutations causales ou à des marqueurs dans des régions à effet quantitatif. Cette caractérisation précoce du génotype d'un individu permet la mise en place de protocoles de sélection plus rapide. La détermination des génotypes de part et d'autre du locus du gène à sélectionner permet de ne conserver que les individus porteurs de l'allèle favorable. Dans le cas d'un schéma d'introgression, la sélection de l'allèle favorable issu d'une population « donneuse » et le retour pour les autres locus aux allèles de la population receveuse peut être également accéléré par l'utilisation de marqueurs moléculaires (Hospital *et al*., 1992).

Cette approche peut être aussi envisagée simultanément sur plusieurs régions QTL dans des protocoles d'évaluation des reproducteurs, grâce à la mise en évidence de marqueurs étroitement liés au QTL.

Cette stratégie a notamment été utilisée dans le cadre d'un programme de sélection assistée par marqueurs entre l'INRA et un organisme de sélection porcine (ADN) pour une lignée composite sino européenne (Duochan) (Schwob *et al*., 2009). La sélection des verrats a été réalisée en 2 étapes. Tout d'abord, les animaux ont été classés en fonction d'un indice simplifié prenant en compte 4 critères : l'âge des animaux, l'épaisseur de lard dorsal, la qualité des côtelettes à 100 kg et la qualité des aplombs. Puis les animaux qui présentaient les meilleurs scores ont été génotypés pour un ensemble de marqueurs encadrant les 4 principales régions QTL mises en évidence dans le programme PORQTL (Bidanel *et al*., 2001). L'obtention des génotypages de chaque individu permet alors d'établir un score moléculaire qui tient compte du poids économique du caractère et de la probabilité du génotype au QTL. Enfin, l'analyse combinée des 2 scores, phénotypique et moléculaire, a permis la sélection des meilleurs individus.

Ce programme est le seul programme français de sélection assistée par marqueurs pour des caractères quantitatifs qui a été mené chez le porc ; cela est certainement dû au fait que l'efficacité de la SAM dépend principalement de la fiabilité de l'estimation de l'effet du génotype au QTL. En 2008, les intervalles de localisations étaient encore trop importants pour que les marqueurs utilisés permettent de sélectionner les allèles favorables aux QTL avec une probabilité forte.

Cependant un programme similaire de plus grande ampleur a été mené dans les 3 principales races bovines laitières françaises. Chaque année 10 000 animaux ont été génotypés pour une quarantaine de marqueurs localisés dans 14 régions chromosomiques. Un premier bilan a permis de mettre en évidence un gain d'efficacité de 5 à 8% selon les caractères par rapport à une sélection classique (Guillaume *et al*., 2008).

2 STRATEGIES DE CARTOGRAPHIE GENIQUE CHEZ LE PORC

Si à partir des années 1990 la recherche de mutation causale est apparue possible, les stratégies mises en œuvre ont fortement évolué au cours du temps. La démarche pour localiser et identifier les gènes impliqués dans la variabilité des caractères quantitatifs a connu un grand changement à partir des années 2000, avec le séquençage des génomes complets des animaux domestiques.

L'objectif de cette seconde partie est donc de faire un court rappel des concepts de base en cartographie génétique, puis de comparer les stratégies de clonage positionnel mises en place avant et après la disponibilité de la séquence. Le sujet d'étude de ce mémoire étant la cartographie fine et la caractérisation d'un QTL influençant l'adiposité des animaux, localisé sur le chromosome 1 porcin, cette partie sera illustrée chaque fois que c'est possible par des exemples de la stratégie de cartographie de ce QTL d'intérêt.

2.1 Concepts de Base et Définitions

2.1.1 Polymorphisme

Un polymorphisme correspond à la variation nucléotidique d'une base, voire d'un segment d'ADN, à un locus donné. Si cette différence induit une modification de l'expression d'un gène, elle peut alors induire des caractéristiques phénotypiques différentes entre plusieurs individus d'une même population. Mais une grande majorité des polymorphismes n'ont pas d'effet phénotypique.

En génétique moléculaire ces variations de l'ADN permettent d'établir l'origine parentale de l'allèle et donc de distinguer les chromosomes issus du père et de la mère. La qualité d'un marqueur génétique peut être évaluée par la mesure du PIC (Polymorphic Information Content (Botstein *et al*., 1980)), qui représente la proportion d'individus informatifs dans une population (Boichard *et al*., 1998). Un marqueur est d'autant plus informatif que le nombre d'allèles est élevé et que leurs fréquences sont équilibrées.

En pratique, le polymorphisme peut être de diverses origines et mis en évidence par un très grand nombre de techniques. Cependant, les marqueurs les plus utilisés en génétique moléculaire sont les microsatellites et les SNP (Single Nucleotide Polymorphim). Les marqueurs microsatellites sont des répétitions simples ou multiples d'un motif court composé de 2 à 10 nucléotides. C'est le nombre variable de répétitions qui induit la variabilité allélique. Pour que ce marqueur puisse être un repère spécifique d'une région du génome, il doit être entouré à droite et à gauche de séquences uniques. Le fréquence des microsatellites dans les génomes de mammifères est d'environ 1/50 000 pb (Schibler *et al*., 2000). En génétique humaine, murine et des animaux domestiques, ce type de marqueurs a été choisi pour développer les premières cartes génétiques denses de référence balisant régulièrement les génomes. Les marqueurs SNP se caractérisent quant à eux par la variation (polymorphisme) d'une seule paire de bases du génome. Les SNP sont très généralement bi-alléliques mais présentent l'avantage d'être très fréquents (1/500 pb) (Vignal *et al*., 2002). Ces polymorphismes ont été peu utilisés initialement comme marqueurs génétiques sur les premières cartes des animaux domestiques, mais depuis 2009 avec les évolutions technologiques ils sont devenus les marqueurs de prédilection pour toute analyse de la variabilité des génomes (Sanchez *et al*., 2012).

2.1.2 La liaison génétique

L'établissement d'une carte génétique est basé sur la notion de liaison génétique et d'une analyse statistique. L'analyse de liaison consiste à observer la transmission allélique de 2 ou plusieurs loci au sein de

familles informatives pour estimer les taux de recombinaison entre ces loci (θ) et tester leur liaison génétique (θ< 0.5). Prenons l'exemple de deux loci M et N (présentant chacun deux allèles 1 et 2) et hétérozygotes chez le parent analysé. Dans l'exemple représenté sur la Figure 5 les allèles M1 et N1 sont portés en phase (c'est-à-dire sur le même chromosome) sur le chromosome paternel, les allèles M2 et N2 en phase sur le chromosome maternel. Au cours de la méiose une copie sera transmise à la descendance, de type parentale (paternelle ou maternelle) ou de type recombinée (résultant du phénomène de recombinaison méiotique). Si l'évènement de recombinaison survient entre les marqueurs M et N, les gamètes recombinés porteront en phase les allèles M1 et N2 ou M2 et N1. En général, plus deux marqueurs sont physiquement éloignés sur le chromosome, plus ils ont de chances d'être séparés par un événement de recombinaison. La mesure du taux de recombinaison reflète donc la distance entre les loci, qui est exprimée en unités appelée "centiMorgan" (1 centiMorgan (cM) correspond à 1% de recombinaison c'est-à-dire une recombinaison en moyenne pour 100 méioses observées).

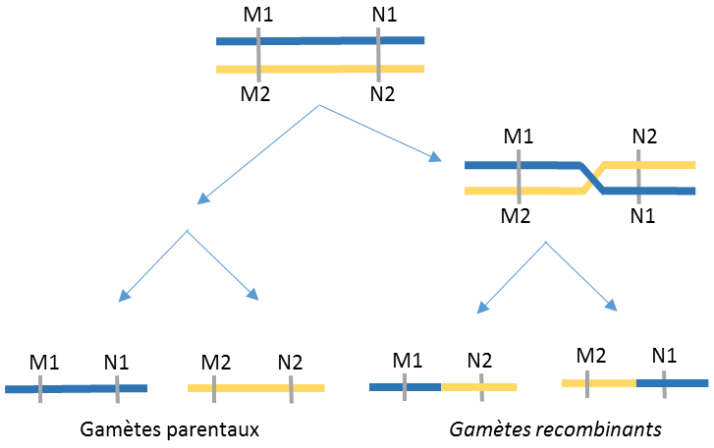


Figure 5 : Illustration de la liaison génétique et de la recombinaison.

Plus ces deux loci sont proches l'un de l'autre, plus la probabilité qu'un évènement méiotique induise une recombinaison de phase sera faible. Si les produits de méiose comportent autant des 4 phases possibles (M1N1, M2N2, M1N2 et M2N1), on conclura à l'indépendance entre les marqueurs : les loci sont portés par deux chromosomes différents ou sont très éloignés sur un même chromosome. La liaison génétique reflète donc une distance entre les loci qui n'est pas totalement dépendante de la distance physique (estimée en nombre de bases). Le ratio cM/pb peut varier d'une région à l'autre au sein du génome d'une espèce, et entre les espèces. Néanmoins chez les mammifères, on admet qu'1 cM correspond à 1 Mb (Mégabase) en moyenne ; chez les oiseaux ce ratio est plus important, avec 2,5 cM pour 1 Mb en moyenne.

2.1.3 Le Déséquilibre de Liaison

2.1.3.1 Définition du déséquilibre de liaison

La notion de déséquilibre de liaison (DL) reflète l'association préférentielle de certains allèles de paires de marqueurs par rapport à l'association théorique attendue dans une distribution aléatoire dans une population.

Le déséquilibre de liaison peut être évalué de la façon suivante :

$$D_{MN} = f_{M1N1} - f(M1) \times f(N1)$$

Où f_{M1N1} est la fréquence observée de l'haplotype (segment chromosomique caractérisé par ses allèles aux variants SNP) porteur des allèles M1 et N1 et f(M1) et f(N1) correspondent aux fréquences des allèles M1 et N1 dans la population. Si la valeur D = 0 il y a équilibre gamétique, si $0 < |D| < 1$, il y a un déséquilibre de liaison entre

les loci M et N. Si |D| = 1 le déséquilibre est total, M1 est toujours associé à N1 et aucun individu ne portera simultanément les allèles M1 et N2.

Cette mesure D du DL dépend des fréquences alléliques et n'est pas normalisée. Elle ne permet pas de comparer des valeurs entre plusieurs paires de locus. C'est pourquoi d'autres mesures ont été proposées comme la valeur de corrélation r^2 par Hill and Roberston (1968) ou le D' par Lewontin (1964).

2.1.3.2 Les facteurs influençant le déséquilibre de liaison

De nombreux facteurs peuvent influencer l'étendue de ces blocs haplotypiques. Certains facteurs vont avoir tendance à augmenter ce DL (les mutations, la sélection, ou les facteurs démographiques) alors que d'autres au contraire vont tendre à le diminuer (la recombinaison).

- Les mutations : La mutation est le phénomène évolutif qui crée un nouveau polymorphisme qui sera en déséquilibre total avec les allèles aux SNP, portés par l'haplotype dans lequel la mutation est apparue. La mutation est donc le moteur de la création du DL. Le DL aux alentours des nouveaux polymorphismes restera important (Figure 6) tant qu'il ne sera pas dissipé par la recombinaison.
- La sélection : Si une mutation confère un avantage sélectif, la sélection de ce locus va entrainer une augmentation locale du DL par effet autostop génétique. Ce phénomène s'accompagnera également d'une baisse de la diversité allélique aux alentours du locus sélectionné (Figure 6).

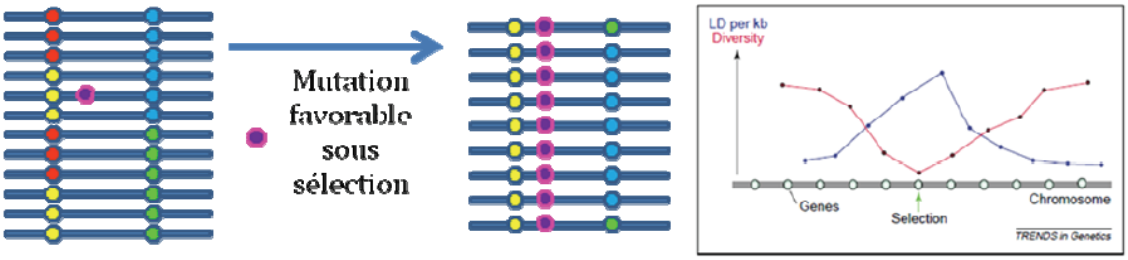


Figure 6 : Mutation entrainant un avantage sélectif.

S'il apparaît dans une population initiale (à gauche) une mutation favorable, celle-ci va peu à peu envahir la population, entrainant avec elle les polymorphismes proches. Adapté de Rafalski & Morgante (2004).

- Les facteurs démographiques : la dérive génétique ou les goulots d'étranglement, vont avoir un impact direct sur la taille efficace des populations et donc sur l'étendue du DL. En effet, les populations de faibles effectifs vont avoir tendance à perdre de manière continue et progressive des allèles rares. Ce phénomène tend à augmenter le DL. De la même manière, un goulot d'étranglement va engendrer de façon temporaire un DL très élevé qui subsistera tant que les évènements de recombinaisons ne le casseront pas.

- La recombinaison : c'est le facteur majeur de perte du déséquilibre de liaison. En l'absence de sélection, la décroissance au cours du temps du DL est fonction du taux de recombinaison. Plus 2 marqueurs seront proches et donc moins le taux de recombinaison sera élevé, et moins le DL diminuera à la génération suivante. Le taux de recombinaison étant variable selon les régions du génome, on peut donc en déduire que le DL sera également variable en fonction des zones étudiées. Des blocs d'haplotypes conservés pourront donc être séparés par des points chauds de recombinaison. Il est à noter que la variation du taux de recombinaison peut être très schématiquement superposée au contenu en gènes de la région considérée (Rafalski and Morgante, 2004).

2.2 Stratégie de cartographie des QTL avant 2009

Les premières cartes du génome du porc ont été développées dans le cadre de consortium, comme PigMap. Trois publications majeures ont permis de disposer de 3 cartes génétiques, composées essentiellement de marqueurs microsatellites et les premiers protocoles de recherche de gènes majeurs et de QTL ont été mis en place. Jusqu'en 2009, la démarche générale pour mettre en évidence un locus influençant un caractère quantitatif (QTL) a principalement reposé sur des **analyses de liaison** au sein de dispositifs familiaux et se décomposait en 2 grandes étapes, la primo-localisation puis la cartographie fine.

2.2.1 Principe de primo localisation de QTL par analyse de liaison

Le principe général de la primo-localisation d'un QTL consiste à observer, dans une famille issue d'un reproducteur hétérozygote pour un marqueur (M1/M2), s'il existe une différence de performance entre les 2 groupes de descendants classés selon l'allèle reçu au marqueur (M) (Figure 7). Si les moyennes des performances des descendants ayant reçu respectivement l'allèle 1 et l'allèle 2 sont significativement différentes on peut en déduire qu'il existe un QTL influençant le caractère proche du marqueur M ; on dit alors qu'un QTL, au voisinage de M, est détecté (Le Roy and Elsen, 2000).

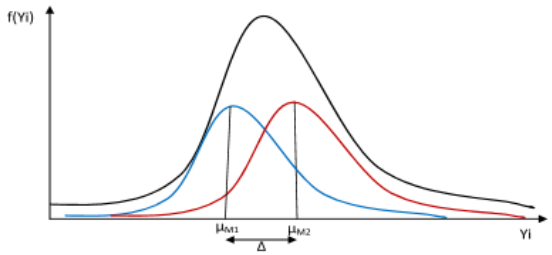


Figure 7 : Courbes de distribution des phénotypes en fonction des allèles au marqueur reçu.
La courbe noire représente la distribution des performances enregistrées sur l'ensemble des individus, la courbe bleue correspond à la distribution phénotypique des descendants qui ont reçu l'allèle M1 au marqueur et la courbe rouge, celle des descendants ayant reçu l'allèle M2. La moyenne générale des performances est μ, celle des descendants qui ont hérité l'allèle M1 au marqueur est μM1 et les individus qui ont reçu l'allèle M2 au marqueur ont μM2 comme moyenne. L'effet de substitution est défini par Δ= μM2 - μM1.

Comme aucune information n'est disponible sur la localisation du QTL recherché, il est nécessaire que cette analyse de co-ségrégation des allèles aux marqueurs et des performances soit menée sur l'ensemble du génome. La stratégie consiste donc à sélectionner et génotyper un ensemble de marqueurs informatifs (hétérozygotes chez les parents), localisés tout au long du génome.

L'approche simple de cartographie de QTL par le test successif de chaque marqueur présente néanmoins quelques inconvénient : (1) l'effet du QTL évalué sur base de la moyenne des performances des descendants ayant reçu l'un ou l'autre allèle au marqueur est très souvent sous-estimé ; (2) la localisation du QTL est peu précise et un QTL à fort effet éloigné du marqueur génétique conduira au même résultat qu'un QTL à petit effet à proximité du marqueur ; (3) le nombre de descendants doit être important pour que la puissance du dispositif soit suffisante. Afin de remédier à ces problèmes, Lander & Botstein ont proposé en 1989, une approche de cartographie d'intervalle, ou "interval mapping", basée sur la prise en compte simultanée de l'information apportée par plusieurs marqueurs (Lander and Botstein, 1989). En effet, lorsqu'une carte génétique est disponible il est possible de considérer les marqueurs 2 à 2 et de définir des intervalles successifs sur l'ensemble du génome et de tester si un QTL est situé entre ces 2 marqueurs.

Le génotypage de 100-150 marqueurs permet dans le dispositif familial de connaitre les allèles transmis des parents aux descendants, aux positions des marqueurs analysés, et de déduire également (aux

recombinaisons près) les phases parentales reçues entre ces marqueurs par chaque descendant. La cartographie d'intervalle permet ainsi une analyse de co-ségrégation entre les performances et une probabilité de génotype à intervalle régulier (1 cM).

Pour un QTL, hétérozygote Q/q (Q correspond à l'allèle qui présente des performances les plus importantes et q correspond à l'allèle avec des performances moindres) chez les parents du dispositif, l'utilisation de l'information de 2 marqueurs encadrant la position du QTL permet de prédire avec une plus grande probabilité la transmission de l'allèle Q, en fonction des haplotypes reçus. Cette approche, plus puissante qu'une approche uni marqueur, permet de réduire le nombre de descendants nécessaire et de discerner mieux la position et l'effet du QTL (Figure 8).

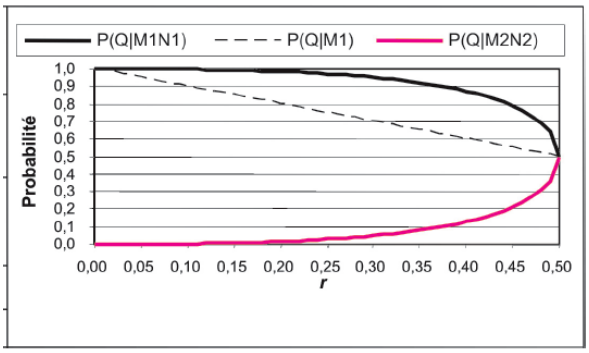


Figure 8 : Probabilités qu'un descendant reçoive l'allèle Q de son père sachant qu'il a reçu M1N1 aux marqueurs flanquant (le QTL étant à équidistance « r », entre les marqueurs M et N).

Les traits pleins correspondent aux probabilités de porter l'allèle Q au QTL en utilisant l'information apportée par 2 marqueurs (transmission des allèles M1 et N1, ou M2 et N2) ; le trait en pointillés est la probabilité de transmission de l'allèle Q si seul le marqueur M est pris en compte.

Afin d'évaluer l'existence d'un QTL en une position du génome, plusieurs méthodes statistiques existent mais la plus courante est le test de rapport de vraisemblance : (1) les probabilités de transmission de phase sont estimées comme présenté ci-dessus, (2) l'effet du QTL supposé à la position testée est estimé par le coefficient de régression linéaire de la performance sur cette probabilité de génotype et (3) les vraisemblances des données sont estimées sous les hypothèses H0 (absence de QTL) et H1 (présence de QTL). Si le rapport des vraisemblances dépasse un certain seuil, l'hypothèse d'absence de QTL est rejetée. La position du QTL la plus probable est celle qui maximise le rapport de vraisemblance.

Lorsque l'existence d'un QTL en une position x est finalement retenue, il est indispensable de définir l'intervalle de confiance dans lequel peut se situer ce QTL. La démarche la plus couramment utilisée consiste à prendre comme borne de l'intervalle les positions entourant la position x auxquelles la probabilité d'existence du QTL est 10 fois moins grande qu'en x *(Ott, 1999)*. Une autre méthode couramment utilisée est de réaliser des simulations par bootstraps *(Visscher et al., 1996)*. Cette méthode consiste à réaliser n tirages avec remise des individus, pour générer des nouveaux jeux de données. On effectue ensuite autant d'analyses de cartographie de QTL que de nombres de jeux de données simulées. A chaque fois on retiendra la position de la valeur du maximum du test statistique. L'intervalle de confiance sera défini sur base des quantiles de la distribution des positions obtenues pour chaque simulation.

2.2.1.1 Les dispositifs expérimentaux

La puissance de détection d'un QTL est directement liée au nombre de parents informatifs dans la population étudiée. Dans le cas d'animaux de laboratoire ou de certaines espèces végétales il est possible d'utiliser des lignées consanguines permettant de créer des parents F1 hétérozygotes pour de très nombreux loci. Or de telle lignées n'existent pas dans les populations animales agronomiques, la stratégie est donc d'utiliser

des populations extrêmes, soit de races très différentes de par leurs performances (exemple du Programme PORQTL), soit des lignées extrêmement divergentes (exemple du Programme de détection QTL dans des lignées divergentes pour la consommation alimentaire résiduelle).

Les protocoles QTL consistent alors à produire des animaux F1 puis des croisements de seconde génération F2 ou BackCross (BC) sur lesquels les caractères sont mesurés ; le choix entre ces 2 dispositifs est souvent dicté par des considérations pratiques.

- Les dispositifs BackCross

Une première possibilité consiste à croiser un individu F1, issu d'un croisement entre 2 races divergentes, sur une des deux races parentales.

Les protocoles BC sont généralement faciles à mettre en place et nécessitent la production d'un plus petit nombre d'individus que les dispositifs F2 (Georges, 2007). Cependant, il est nécessaire de connaitre les caractéristiques du caractère étudié (dominance, codominance ou récessivité) afin d'orienter le sens du croisement en retour. En effet, si l'allèle Q identifié présente un effet de dominance, il faudra obligatoirement réaliser un croisement du parent F1 (Qq) avec des animaux homozygotes récessifs (qq) pour pouvoir observer dans la descendance une différence phénotypique entre les 2 classes de génotypes (Qq et qq).

Dans le cas d'une dominance/récessivité clairement identifiée entre les 2 allèles du QTL, le protocole BC s'avère alors plus puissant qu'un protocole F2 pour détecter des QTL (Darvasi, 1998).

- Les dispositifs F2

Le dispositif F2 repose sur le croisement entre 2 individus F1 hétérozygotes, issus également de 2 races très éloignées afin de maximiser les différences phénotypiques au sein de la descendance. L'étude d'une population F2 permet d'obtenir une meilleure estimation du nombre de QTL en ségrégation et d'avoir une « vue d'ensemble » du génome (Darvasi, 1998) quels que soient les effets des QTL. En effet, dans ce type de dispositif les 3 génotypes au QTL (qq, Qq et QQ) peuvent être observés.

2.2.1.2 Un exemple de dispositif expérimental porc : le programme PORQTL

Entre 1992 et 1998, l'INRA a conduit son propre dispositif de détection de QTL (= Programme PORQTL) à partir de croisements expérimentaux de type F2.

Ce programme avait pour objectif d'identifier les régions chromosomiques influençant des caractères de croissance, d'engraissement, de composition de carcasse, de reproduction et de comportement. Afin de maximiser la variabilité des performances, les animaux utilisés pour constituer les familles étaient des verrats de race Européenne (Large White, « Lw ») et des truies de race chinoise (Meishan, « Ms »). En effet, ces 2 races présentent des phénotypes distincts, la race Large White présente de bonnes aptitudes de croissance et de composition de carcasse, alors que la race Meishan présente de bonnes aptitudes maternelles (Milan *et al*., 2002).

Pour détecter et localiser les QTL, 6 verrats Large White et 6 truies Meishan F0 ont été croisés afin d'obtenir des descendants F1. Parmi ces animaux, 6 mâles F1 et 23 femelles F1 ont été conservés pour produire la génération F2. Au total 530 males et 573 Femelles F2 ont été phénotypés (Tableau 1) et génotypés avec 137 marqueurs microsatellites, répartis sur les 18 autosomes et le chromosome X (Figure 9). Le nombre de marqueurs par chromosome varie de 3 (chromosome 18) à 10 (chromosome 7).

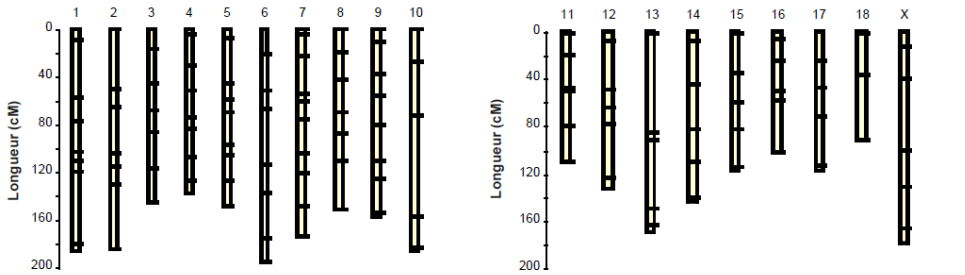


Figure 9 : Répartition des 137 marqueurs microsatellites sur les 18 autosomes et le chromosome X porcin.

Chaque rectangle représente un chromosome et les petites barres horizontales, la position d'un marqueur. La couverture en marqueurs est globalement satisfaisante et bien répartie pour l'ensemble des chromosomes.

La mise en place d'un tel dispositif est extrêmement coûteuse car le fait que ces animaux soient 50 % Meishan fait perdre une importante partie de la valeur commerciale des animaux. Un très grand nombre de caractères ont donc été mesurés pour tirer parti au maximum de ce protocole. Au total, 92 phénotypes ont été mesurés dont les principaux sont mentionnés dans le Tableau1 ((Bidanel *et al*., 2001) ; (Milan *et al*., 2002)).

Tableau 1 : Principaux caractères étudiés dans le projet PORQTL.

Fonction étudiée	Sexe	Caracteres mesurés
Croissance	M	Poids corporels à J0, J21, J100, J140 et J160
	M	Teneur en Hormone de croissance et glycémie avant et apres test à l'insuline (J120)
	F	Poids corporels à J0, J21, J70, J90, J120 et J150
Composition corporelle	M	Epaisseur de lard dorsal à J100, J120, J140 et J160
	M	Rendement de carcasse et poids des morceaux de découpe à 80Kg
	F	Epaisseur de lard dorsal à J90, J120 et J150
Caracteristique du muscle et du gras	M	Taux d'androstenone du gras à J100, J120, J140, J160 et à 80Kg
	M	Taux de scatole et d'indole du gras
	M	Nombre de fibres musculaires % de la partie blanche (long dorsal)
	M	Taux de gras intramusculaire, activité des enzyme de la lipogénèse
Reproduction	M	Longueur et poids des glandes de Cowper, poids des testicules, des épididymes et des vésicules séminales
	F	Age au premier oestrus (examen visuel) et à la puberté (progesterone), taux d'ovulation, nombres d'embryons, poids du tractus génital, des cornes utérines (+ longueur) et des ovaires
Réactivité Neurodocrinienne et Comportementale	M +F	Réponse neuroendocrinienne (cortisol, ACTH, glycémie) et comportementale (locomotion, vocalise, défécations) à un test d'exposition à un environnement nouveau (J42).
Caracteristiques sanguines	M +F	Nombres de globules rouges et de plaquettes, taux d'hémoglobine, volume des plaquettes, dosage du fer total, capacité de saturation en fer et transferrine.

Les analyses de liaison génotype/phénotype ont alors été réalisées à l'aide du logiciel QTLmap (Filangi *et al.,* 2010) afin de primo-localiser les régions du génome contribuant à la variabilité observée au sein des descendants F2.

Deux méthodes statistiques distinctes ont été utilisées (Bidanel *et al*., 2001). La première méthode pose comme hypothèse que chaque population est fixée pour un allèle diffèrent au QTL. L'analyse correspond à l'estimation d'un contraste entre les effets des allèles des 2 races grand-parentales transmis aux animaux F2. Cette stratégie est très puissante si les allèles sont effectivement fixés comme c'est le cas dans des lignées de souris, mais cette hypothèse n'est pas toujours vérifiée dans les espèces d'élevage telles que le porc ou le bovin.

La seconde méthode est plus proche de la réalité, en estimant le contraste entre les allèles parentaux à l'aide des modèles demi-frères/pleins frères. Ce modèle estime, pour un lot de descendants issus d'un même

père, le contraste de l'effet des allèles paternels transmis à sa descendance (half-sib) et les effets de substitution des allèles transmis par la mère au sein d'une fratrie (full-sib).

Ces analyses ont permis de mettre en évidence un très grand nombre de QTL sur l'ensemble des chromosomes à l'exception des chromosomes 10, 12, 15 et 18. Au final, plus de 85 QTL, fortement significatifs avec un seuil supérieur à 0,1 % au niveau du génome ont ainsi été primo localisés (Bidanel *et al*., 2000).

Neuf régions chromosomiques comprenant des QTL liés aux caractères de croissance ont été mises en évidence ; parmi ces régions, 2 sont nettement associées à des variations importantes pour la plupart des caractères de croissance et sont situées sur les chromosomes 4 et 7. L'allèle d'origine Meishan est souvent associé à une moindre vitesse de croissance. Ce résultat est en accord avec ceux obtenus par Andersson *et al*., 1994 et Marklund *et al*., 1999. Les QTL présentant les effets les plus forts pour les caractères de carcasse sont localisés sur les chromosomes 7 et X. Les allèles Meishan localisés sur le chromosome X sont associés à la fois à une augmentation de la quantité de gras et une diminution de la quantité de muscle. Les allèles Meishan du chromosome 7 sont quant à eux associés à une diminution du poids des pièces à l'abattage. Des résultats similaires ont été obtenus par (Rohrer and Keele, 1998). Un des résultats les plus marquants concerne le chromosome 2, pour lequel l'effet favorable porté par un allèle Lw, qui favorise la croissance musculaire (augmentation du poids de la longe) correspond à la région du gène IGF2. Ce gène est très bien documenté dans la littérature et correspond à un facteur de croissance important chez le fœtus et l'adulte, des effets similaires ont été également démontrés dans un croisement Large White x Piétrain (Nezer *et al*., 1999) et un croisement Large White x Sanglier (Jeon *et al*., 1999).

Concernant le caractère d'épaisseur de lard dorsal, 5 chromosomes atteignent le seuil de significativité de 0,1% au niveau du génome (Tableau 2). Les régions détectées qui présentent les plus forts effets sont localisées sur les chromosomes 1, 4, 5, 7 et X (Bidanel *et al*., 2001). Les régions sur les chromosomes 1 et 7 sont les mêmes que celles identifiées pour des caractères de croissance. Les allèles Meishan des chromosomes 1 et X sont associés à une augmentation de l'adiposité. Inversement, les allèles Meishan du chromosome 7 présentent une moindre épaisseur de lard dorsal.

Tableau 2 : Principaux QTL détectés pour le caractère d'épaisseur de lard dorsal (BackFat).

Tableau issu de l'article Bidanel et al., 2001.

		Line-cross model			Half-/full-sib model		
Trait [a]	SSC	Position and confidence interval (cM)	F-ratio	Significance level [b]	Position and confidence interval (cM)	LR [c]	Significance level [b]
BF13w	1	175 (167–175)	43.2	***	172 (162–175)	156.3	***
BF17w	1	175 (167–175)	36.6	***	175 (163–175)	133.1	***
BF22w	1	175 (161–175)	12.6	***	172 (161–175)	82.7	***
BF40kg	1	175 (166–175)	31.4	***	171 (161–175)	108.5	***
BF60kg	1	175 (165–175)	24.3	***	175 (163–175)	99.8	***
BF40kg	4	62 (49–84)	15.3	***	62 (55–82)	88.8	***
BF60kg	4	72 (59–84)	14.9	***	63 (55–84)	70.6	*
BF13w	5	35 (20–46)	8.3	+	– [d]	–	ns
BF17w	5	41 (33–48)	15.1	***	43 (34–49)	60.8	+
BF40kg	5	37 (26–45)	13.4	***	43 (33–50)	58.6	+
BF60kg	5	42 (36–49)	23.2	***	43 (36–48)	87.1	***
BF13w	7	65 (57–72)	29.9	***	69 (56–80)	90.4	***
BF17w	7	58 (51–68)	30.0	***	73 (63–84)	94.6	***
BF22w	7	57 (51–67)	35.6	***	73 (63–84)	92.2	***
BF40kg	7	65 (57–71)	75.7	***	67 (61–73)	182.6	***
BF60kg	7	62 (55–68)	87.5	***	66 (58–72)	189.8	***
BF13w [e]	X	74 (69–84)	31.8	***	80 (71–88)	112.6	***
BF17w [e]	X	73 (68–85)	31.6	***	78 (69–89)	91.6	***
BF22w	X	73 (65–84)	27.8	***	77 (70–87)	96.3	***
BF40kg [e]	X	74 (68–85)	45.3	***	83 (76–90)	155.1	***
BF60kg [e]	X	74 (69–84)	35.7	***	81 (68–89)	119.2	***

[a] See Table II for the definition of the traits. [b] *; **; *** = 5%, 1% and 0.1% genome-wide significance levels, respectively. + = suggestive linkage. [c] LR = likelihood ratio. [d] Not estimated. [e] In males only.

Ces résultats sont également en accord avec ceux précédemment obtenus par d'autres équipes. Pour exemple, un QTL d'épaisseur de lard dorsal à bien été retrouvé localisé à l'extrémité basse du chromosome 1 entre les positions 120 et 160 cM (Figure 10, Rohrer and Keele, 1998)

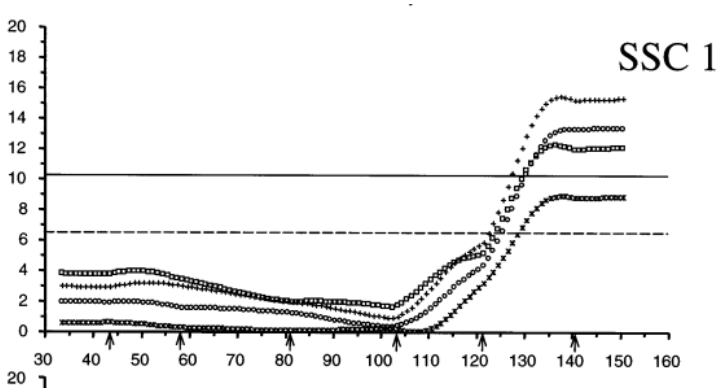


Figure 10 : Profil de la statistique de test pour l'épaisseur de lard dorsal (ELD) sur le chromosome 1.

L'axe des abscisses représente la position relative sur le chromosome 1 en cM et les flèches indiquent les positions des marqueurs microsatellites utilisés pour le test statistique. L'axe des Y représente la valeur du test de Fisher. Les lignes horizontales correspondent au niveau de significativité de ce test au niveau du génome, une valeur dépassant le trait plein (> 10.15) est considérée comme significative et une valeur comprise entre 6.6 et 10.15 est considérée comme suggestive. Les 4 courbes représentent les profils de statistiques pour les mesures de caractères d'épaisseurs de lard dorsal, o = mesure d'épaisseur de lard dorsal à la dernière côte, x = mesure d'épaisseur de lard dorsal à la 20ieme côte, □ = mesure d'épaisseur de lard dorsal à la dernière lombaire, + = moyenne des mesures d'épaisseur de lard dorsal (Rohrer and Keele, 1998).

Pour les caractères de reproduction, peu de QTL significatifs ont été trouvés. Le résultat le plus net porte sur la longueur ou le poids des cornes utérines. Les QTL qui avaient été mis en évidence sur le taux d'ovulation et le nombre d'embryons lors de résultats préliminaires (Bidanel et al, 1998) ne sont plus détectés malgré un nombre plus important d'individus. Ces résultats peuvent s'expliquer par le fait que probablement plusieurs QTL sont à l'origine de la plus grande prolificité de la race Meishan et que le dispositif expérimental utilisé reste trop peu puissant pour des caractères qui présentent une si faible héritabilité.

Comme précédemment évoqué, plusieurs programmes similaires ont été initiés par différentes équipes étrangères. L'ensemble des données relatives à la localisation des différents QTL détectés sont stockés dans une base de données publique appelée PigQTLdb http://www.animalgenome.org/cgi-bin/QTLdb/SS/index.

Ainsi à partir de tous ces programmes, 7169 QTL ont été détectés par analyses de liaison et sont référencés dans 414 publications. Ces QTL sont répartis sur tous les chromosomes porcins et affectent 554 caractères différents. Ce nombre de QTL est largement surestimé, car les QTL mis en évidence dans des protocoles indépendants mais influençant un même caractère et localisés dans un intervalle similaire peuvent très certainement correspondre à un même locus.

Généralement à l'issue de cette première étape de primo-localisation, les QTL sont localisés dans des intervalles de l'ordre de 20 à 40 cM ; ces intervalles correspondent chez les mammifères à environ 20 à 40 mégabases et comprennent de 200 à 400 gènes.

Il était donc nécessaire dans un second temps de mener une approche dite de cartographie fine afin de réduire cet intervalle. Mais pour pouvoir mener à bien cette seconde étape il est nécessaire :

- D'une part de posséder des outils de cartographie régionale pour augmenter la densité en marqueurs informatifs dans la région d'intérêt afin d'affiner l'intervalle de localisation.
- D'autre part, de pouvoir ajouter au dispositif initial des nouveaux individus porteurs de recombinaisons dans l'intervalle.

2.2.2 Outils de cartographie physique régionale chez le porc

En parallèle des cartes génétiques permettant les analyses de ségrégation, les programmes de cartographie des génomes avaient comme objectif d'affiner la caractérisation des chromosomes (identification des gènes, des éléments de régulation…). Pour cela, divers outils de cartographie avec des niveaux de résolution croissante ont été développés afin d'améliorer la connaissance précise d'un génome en utilisant parfois les données d'espèces mieux cartographiées. Ces approches font appel à plusieurs stratégies de cartographie et aboutissent à quatre types de cartes différentes (Figure 11).

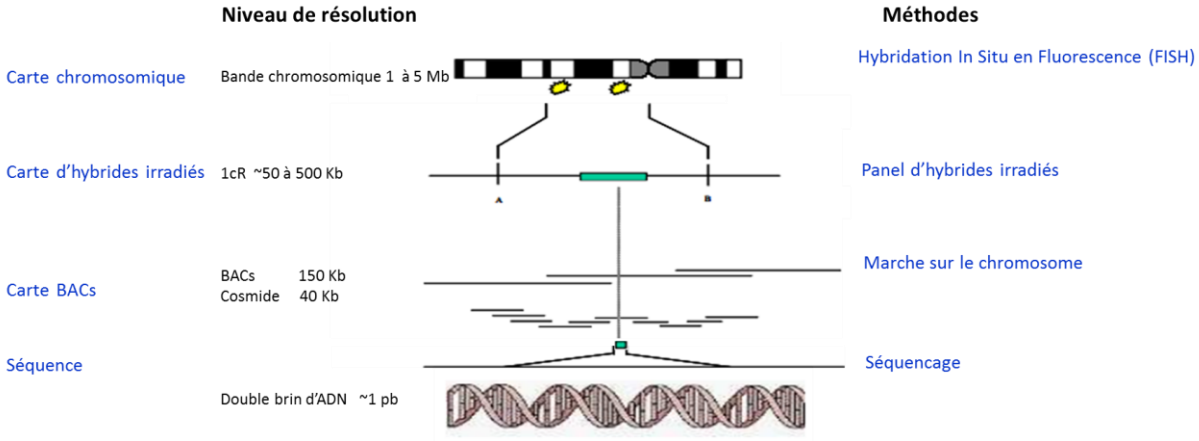


Figure 11 : Les différents niveaux de résolution des cartes et les techniques associées.

2.2.2.1 La cartographie chromosomique hétérologue

La cartographie chromosomique comparée a pour objectif d'identifier les régions chromosomiques synténiques entre espèces. L'identification de ces régions conservées permet en effet de transposer les données acquises chez des espèces modèles très étudiées comme l'homme, la souris, le rat, ou le poisson zèbre, à d'autres espèces animales. Les résultats de la cartographie comparée représentent un atout à l'identification de gènes en tirant parti des informations acquises chez les génomes modèles et apportent des connaissances concernant l'évolution des génomes. Les premières techniques utilisées ont été les hybridations de sondes radioactives puis fluorescentes correspondant à une région de localisation connue chez l'homme (le plus souvent un gène) sur des chromosomes en métaphase de l'autre espèce. Par la suite une approche systématique appelée Zoo-Fish ou coloriage chromosomique a été menée : cette technique correspond à l'hybridation d'une sonde fluorescente complexe (= mélange de fragments spécifiques et représentatifs d'un chromosome) d'une espèce sur les chromosomes en métaphase d'une autre espèce. Cette méthode a permis de déterminer pour l'ensemble du génome les homologies entre les cartes cytogénétiques humaine et porcine (Goureau *et al*., 1996).

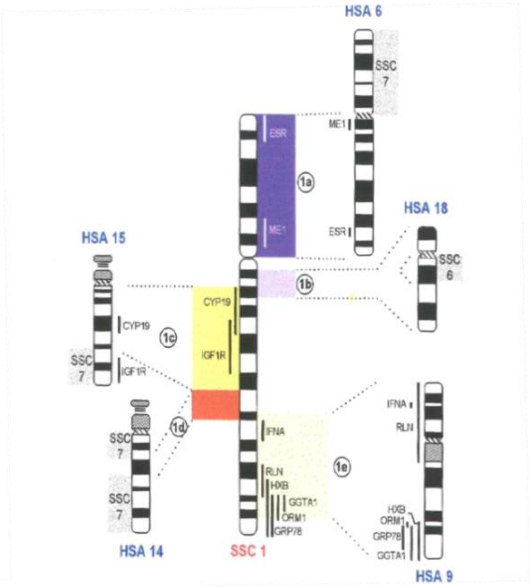


Figure 12 : Correspondance chromosomique entre le porc (SSC1) et l'homme par coloriage chromosomique (Goureau, 1997).

Par Zoo-FISH, Le chromosome 1 porcin (SSC1) est colorié par 5 sondes de chromosomes humains, HSA 6, 9, 14, 15 et 18 (Figure 12). Cette approche a permis de déterminer que la région du QTL d'épaisseur de lard dorsal localisée en 1q24-q212 (zone 1e sur la Figure 12) était homologue à la région 9q32-q34 humaine. Néanmoins la résolution de cartographie obtenue par cette approche est relativement faible, de l'ordre de 5Mb.

2.2.2.2 Cartographie par hybrides d'irradiation

Afin d'obtenir un niveau de résolution plus important, des panels d'hybrides d'irradiation ont été développés afin de construire des cartes d'irradiation. Ces hybrides sont obtenus suite à la fusion d'une cellule donneuse soumise préalablement à une radiation X pour en fragmenter les chromosomes, avec une cellule de Hamster receveuse. Lors de la culture des cellules hybrides (Porc/Hamster), on observe une perte aléatoire des fragments de chromosomes de la lignée donneuse alors que les chromosomes de la lignée receveuse sont conservés. Une collection de clones de cellules hybrides contenant chacun des fragments différents des chromosomes de l'espèce donneuse sera nécessaire pour couvrir l'ensemble du génome (Figure 13).

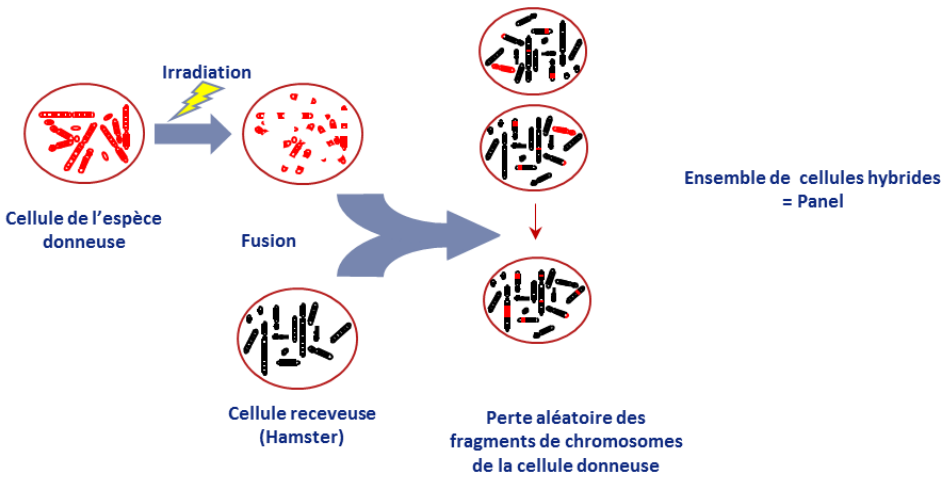


Figure 13 : Construction d'un panel d'hybrides d'irradiation.

Le principe de la cartographie par hybride d'irradiation (RH), illustré dans la Figure 14, repose sur l'analyse de la fréquence de cassures entre les marqueurs, induite par l'irradiation du génome (Barrett *et al*., 1992). En effet plus deux marqueurs seront proches sur le génome, plus leurs distributions au sein des clones hybrides seront semblables car la probabilité qu'ils soient séparés par une cassure liée à l'irradiation sera faible. Un des avantages de la méthode RH par rapport à une approche cytogénétique réside dans la densité plus élevée de marqueurs qui peuvent être analysés. De plus, contrairement aux cartes de liaison génétique qui nécessitent l'utilisation de marqueurs polymorphes et informatifs pour distinguer les 2 allèles transmis, la cartographie RH teste la présence ou l'absence du marqueur dans les différentes lignées du panel, et ne nécessite pas de polymorphisme. Elle permet donc d'utiliser un plus grand nombre de marqueurs et en particulier des marqueurs définis dans la séquence codante, de type EST (Expressed Sequence Tag), souvent non polymorphes.

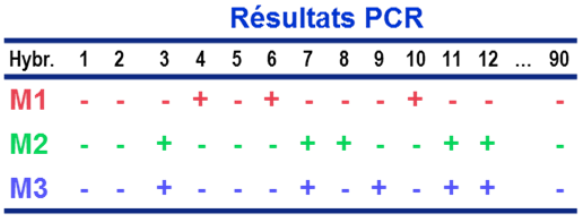


Figure 14 : Génotypage et construction des cartes.

La présence ou l'absence des fragments chromosomiques correspondant aux marqueurs M est testée par PCR sur 90 clones différents, qui constituent le panel. Les marqueurs M2 et M3 ont des profils très semblables, indiquant qu'ils sont proches. M1 est soit plus loin sur le même chromosome, soit sur un autre chromosome.

Deux panels d'hybrides d'irradiation porcins ont été construits à l'INRA de Toulouse (Yerle *et al*., 1998; Yerle *et al*., 2002). Le premier panel, construit en 1998, nommé IMpRH est issu d'une irradiation à 7000 rad. Après la caractérisation de chaque clone cellulaire (118 hybrides) à l'aide d'un très grand nombre de marqueurs, des cartes de référence pour chacun des chromosomes porcins ont été réalisées. A l'aide de ce panel des marqueurs distant de 50 kb ont pu être ordonnés les uns par rapport aux autres (1 $cR_{7000rad}$ / 50 kb).

Le second panel d'hybrides d'irradiation (IMpRH2), développé en 2002, est issu d'une irradiation à une 12000 rad, afin d'obtenir une résolution plus fine encore (1$cR_{12000rad}$ correspond à 20 kb).

Les panels d'hybrides d'irradiation ont été largement utilisés pour préciser les cartes comparées entre espèces. En effet, ils offrent une résolution plus fine (quelques kb) que les approches de cartographie chromosomique hétérologues (5 à 10 Mb) et sont plus faciles d'utilisation car une simple PCR suffit.

De cette manière, une carte d'irradiation a été entreprise pour la région du QTL du chromosome 1 afin de valider et de préciser l'homologie de cette région porcine avec la région humaine HSA 9q32-q34.

Pour cela, des amorces ont été choisies dans les séquences codantes (EST) porcines homologues de chaque gène humain annoté dans la région 9q32-q34 afin de les localiser chez le porc à l'aide du panel IMpRH (Figure 15).

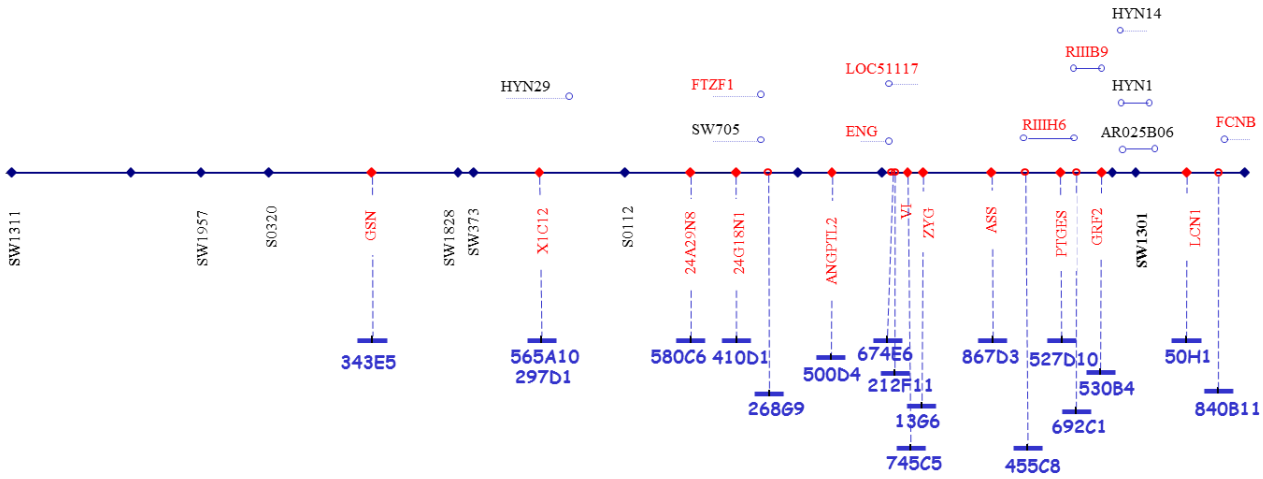


Figure 15 : Position des marqueurs développés sur la carte d'hybrides d'irradiation.

En noir sont indiqués les marqueurs microsatellites de la carte génétique porcine, en rouge les EST (sélectionnés sur base des données de cartographie comparée avec l'homme) et en bleu, les clones BAC qui ont été criblés à partir des EST.

2.2.2.3 La cartographie physique par BAC

Au fur et à mesure du temps, de nouvelles technologies ont été mises en place afin de parvenir à une résolution de cartographie de plus en plus fine. Des résolutions de quelques dizaines de kb ont été obtenues à l'aide de banques de grands fragments d'ADN comme les banques BAC (Bacterial Artificial Chromosome). Des banques de BAC ont été construites chez de nombreuses espèces telles que l'homme (Asakawa *et al*., 1997), le bovin (Cai *et al*., 1995), la chèvre (Schibler *et al*., 1998) et le porc (Suzuki *et al*., 2000). Chez le porc, il existe en tout 5 banques de BAC : les banques CHORI-242 (Osoegawa *et al*., 1998), RPCI44 (Fahrenkrug *et al*., 2001), PigE (Anderson *et al*., 2000), INRA (Rogel-Gaillard *et al*., 1999) et KNP (Jeon *et al*., 2003).

La banque de BAC porcine développée par l'INRA comporte 107 520 clones, dont les insertions présentent une taille moyenne de 135 kb ; l'ensemble des clones de la banque permet une couverture moyenne du génome équivalente à 5x (Rogel-Gaillard *et al*., 1999).

Les banques BAC ont été également très utilisées pour l'obtention de la séquence de référence du génome du porc. En effet, la stratégie qui a été choisie pour le génome de cette espèce est la réalisation du séquençage après un ordonnancement hiérarchique des clones BAC pour définir le nombre minimum de fragments nécessaires permettant de couvrir l'ensemble du génome (le Minimum Tiling Path) (Humphray *et al*., 2007).

Chaque type de carte présente des avantages et des inconvénients. La carte chromosomique classique a une faible résolution et ne permet pas de préciser l'ordre des gènes, cependant c'est la seule carte qui permet d'établir un lien direct avec les chromosomes en conformation métaphasique. A l'inverse les cartes génétiques permettent d'ordonner des marqueurs génétiques au sein d'un groupe de liaison, mais un groupe de liaison sans ancrage chromosomique ne permet pas de savoir où ces marqueurs sont localisés dans le génome et si le groupe de liaison couvre la totalité du chromosome. Par ailleurs, les distances données par la carte génétique sont relatives, car elles dépendent du taux de recombinaison qui est variable tout au long du génome. La distance réelle entre les gènes est obtenue par cartographie physique (contig de BAC, voire séquence). Ces différentes cartes sont complémentaires et bien souvent chacune sert à enrichir les autres.

Ces outils de cartographie physique ont été largement utilisés jusqu'en 2009, car ils étaient une aide importante pour localiser et ordonner les gènes tant que nous ne disposions pas de la séquence. Ils permettaient d'accumuler des informations de cartographie à façon dans n'importe quelle région du génome.

2.2.3 Cartographie fine génétique

La cartographie fine a pour principal objectif de réduire la taille de l'intervalle de localisation d'un QTL. Cette approche est basée sur l'exploitation et la caractérisation des évènements de recombinaison et elle peut être conduite suivant 2 approches complémentaires :

- Développer des marqueurs, pour mieux caractériser ces évènements de recombinaison
- Ajouter au dispositif initial des individus porteurs de nouvelles recombinaisons.

Alors que l'étape initiale de primo-localisation est réalisée sur l'ensemble du génome, les travaux de cartographie fine sont menés de façon indépendante pour chacune des régions. En effet, chaque région nécessite la production spécifique d'animaux et le développement particulier de marqueurs. Parmi les régions primo-détectées dans le cadre du dispositif PORQTL, la cartographie génétique fine de quelques QTL a été entreprise, dont le QTL localisé à l'extrémité du chromosome 1 (130-160 cM).

2.2.3.1 Développement de nouveaux marqueurs

Afin de définir plus précisément les points de recombinaison et pour sélectionner de nouveaux verrats recombinants, il était nécessaire de développer de nouveaux marqueurs génétiques. Si aujourd'hui cette étape peut sembler triviale grâce aux séquences des génomes de référence désormais disponibles, jusqu'en 2009 elle représentait une part importante du travail de cartographie des QTL.

Pour la cartographie fine du QTL du chromosome 1, la démarche générale qui a été mise en place est la suivante : à partir des séquences humaines des gènes de la région orthologue de la région porcine d'intérêt, des EST porcines ont été recherchées dans les bases de données. Ces EST porcines ont été localisées sur la carte d'irradiation afin de contrôler leur localisation, avant d'être utilisées pour rechercher des clones BAC de la région d'intérêt. Enfin, à partir de ces clones des nouveaux marqueurs ont été développés soit par sous-clonage puis criblage afin de trouver des marqueurs de types microsatellite, soit par séquençage des extrémités pour rechercher de nouveaux polymorphismes de type SNP.

2.2.3.2 Création d'individus recombinants : croisement en retour ou Backcross

Le principe de cartographie fine par générations successives d'individus recombinants, décrit en Figure 16, repose sur la production de verrats recombinants dans la région du QTL et de leur testage sur descendance afin de déterminer leur statut (homozygote qq ou hétérozygote qQ) au QTL. Pour cela un mâle F1 Hétérozygote au QTL (q/Q) est croisé en retour sur l'une des populations parentales (LwLw). Parmi les descendants BC1 produits, les animaux ayant reçu de leur père un chromosome recombinant (rec) dans la région du QTL sont conservés, le chromosome maternel étant identique (Lw) pour tous les descendants. Ces verrats sont à leur tour croisés avec des femelles Lw afin de produire une centaine de descendants BC2. Les descendants obtenus sont alors classés en fonction du chromosome qu'ils ont reçu de leur père BC1 et les moyennes des performances sont à nouveau comparées entre les 2 groupes de descendants.

Si les moyennes des performances sont significativement différentes entre les deux groupes on peut en conclure que le père BC1 testé est hétérozygote au QTL. Au contraire si les performances sont similaires alors le père testé est homozygote au QTL et le QTL se situe dans la région où les marqueurs portés par les 2 chromosomes sont d'origine Lw (homozygote Lw).

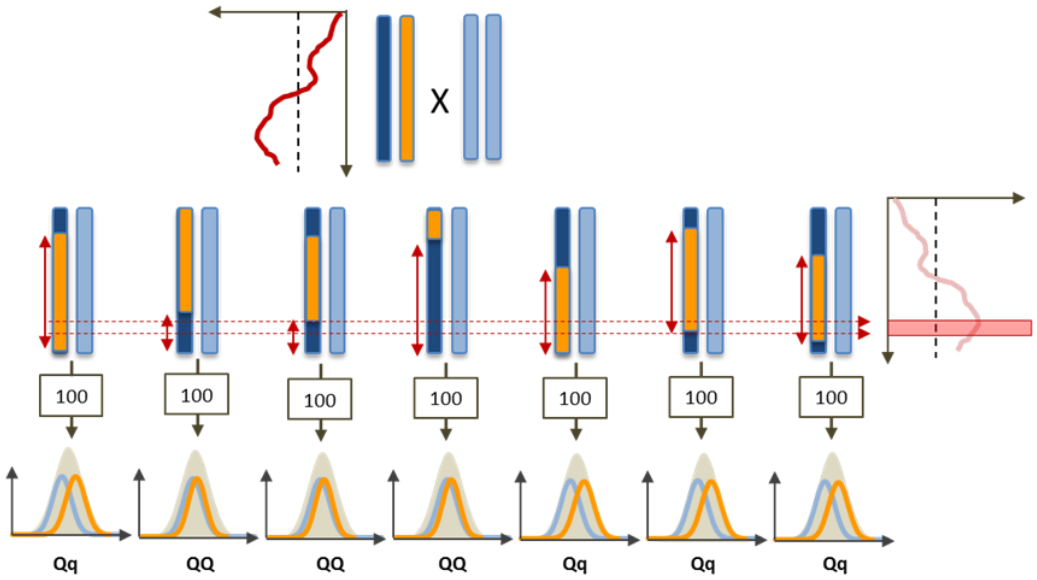


Figure 16 : Stratégie de cartographie fine de QTL à l'aide de croisements en retour (Back-Cross) et testage sur descendance des verrats recombinants.

L'origine raciale des allèles est représentée en bleu (origine Lw) et en jaune (Ms). Le statut au QTL des verrats testés est indiqué en dessous des courbes de distribution des moyennes des performances des 2 classes d'individus (Rec/Lw vs Lw/Lw) (qQ ou QQ). L'intervalle de localisation du QTL déduit pour chaque individu testé est indiqué par une flèche rouge. L'intervalle minimum de localisation du QTL est compris entre les deux traits en pointillés.

Plus le nombre de verrats testés sur descendance présentant des points de recombinaisons différents dans l'intervalle de localisation initial du QTL sera important, plus la taille de l'intervalle déduit suite à leur testage pourra être réduite (Figure 16) (Sanchez *et al.*, 2006). Afin de réduire progressivement l'intervalle de localisation d'un QTL, plusieurs générations d'individus BC recombinants sont souvent nécessaires. Cette stratégie a par exemple été menée avec succès par Berg *et al.*, 2006. En effet 7 générations de BC ont permis de réduire l'intervalle du QTL pour le caractère d'épaisseur de lard dorsal localisé sur le chromosome 4 de 70 cM à 3,3 cM.

La production de familles complémentaires de type BC suite à la primo détection de QTL présente également l'avantage de confirmer ou d'infirmer les résultats obtenus à partir d'un dispositif F2. Ainsi les résultats PORQTL ont été confirmés pour les 4 principales régions QTL en créant des familles backcross (BC) à partir de 2 verrats F1 (Sanchez *et al.*, 2005). Une première génération de BC1 a été réalisée sur des Truies Lw/Lw. Par la suite 4 mâles BC1, hétérozygotes dans 1 à 4 des régions QTL, ont été choisis afin d'être testés sur descendance. Pour cela chaque verrats BC1 a été à nouveau croisé avec une dizaine de truies Lw/Lw. Cette approche est généralement longue et coûteuse. En effet, on ne connait le génotype au QTL des animaux qu'après son testage sur descendance, soit environ 2 ans après sa naissance.

Les résultats du testage de cette génération d'animaux BC1 ont bien confirmé les effets précédemment observés sur les chromosomes 2, 4 et 7.

Mais *a contrario* certains effets n'ont pas été retrouvés, ou avec des effets nettement moins significatifs dans ces analyses que lors des analyses de primo-détection. Les effets très significatifs qui avaient été observés sur l'épaisseur de lard dorsal à l'extrémité terminale du chromosome 1 (Bidanel *et al.*, 2001) n'atteignent ici un seuil de significativité que de 5% à l'échelle du chromosome.

Cela peut s'expliquer par différentes hypothèses :

- **Le nombre d'animaux mesurés** : pour le dispositif BC ce nombre est plus réduit, les effets ont été estimés indépendamment pour chaque famille qui était composée de 29 à 71 descendants alors que les effectifs du programme PORQTL étaient de 120 à 294 descendants par famille.

- **Variabilité allélique :** la détection de QTL repose sur le principe qu'un allèle différent aux QTL est fixé dans chacune des races, ce qui est généralement vérifié par l'absence d'effet pour les verrats homozygotes. Néanmoins, il n'est pas possible d'exclure que plusieurs allèles aux QTL sont en ségrégation dans la population Lw, l'analyse de plusieurs verrats hétérozygotes dans la région du QTL peut alors montrer des résultats différents.
- **Le type génétique des animaux** : comme les effets des QTL sont estimés par comparaison des moyennes des performances des descendants Lw/Ms et Lw/Lw, si l'allèle Lw est partiellement ou totalement dominant sur l'allèle Ms, le sens du croisement en retour vers la race Lw va masquer une partie des effets.

2.2.3.3 Création d'individus BC localement congéniques

Pour valider l'hypothèse d'une dominance partielle de l'allèle de maigre (Lw) par rapport à l'allèle de gras (Ms) et afin de comprendre la perte des résultats significatifs du QTL localisé sur le Chr1, un croisement en retour vers la race Ms aurait pu être réalisé. Cependant compte tenu des coûts de production des animaux Ms ou 1/2Ms (peu valorisables commercialement), il n'était pas possible d'envisager une telle approche.

Pour faire face à cette difficulté, un troupeau de femelles dédiées au testage de cette région a été constitué, avec comme particularités d'être homozygotes Ms/Ms à l'extrémité du chromosome 1 et avec une forte proportion Lw dans le reste de leur génome (29/32 Lw) (Riquet *et al.*, 2011a)

Ces femelles localement congéniques (appelées MsMs dans la suite de ce texte) ont ensuite servi comme support pour réaliser le testage sur descendance des pères recombinants, de la même façon que décrit dans le paragraphe précédent.

Dans un premier temps, afin de valider ce type d'approche, un seul verrat déjà testé à l'aide de femelles supports LwLw, a été testé de nouveau à l'aide de femelles MsMs. Alors que les premiers résultats de ce verrat n'avait pas permis d'obtenir de résultat significatif (Figure 17), ce nouveau dispositif a permis d'obtenir des résultats significatifs au seuil de 0,1% au niveau du génome pour le caractère d'épaisseur de lard dorsal, comme ceux mis en évidence dans la population F2 (Figure 17).

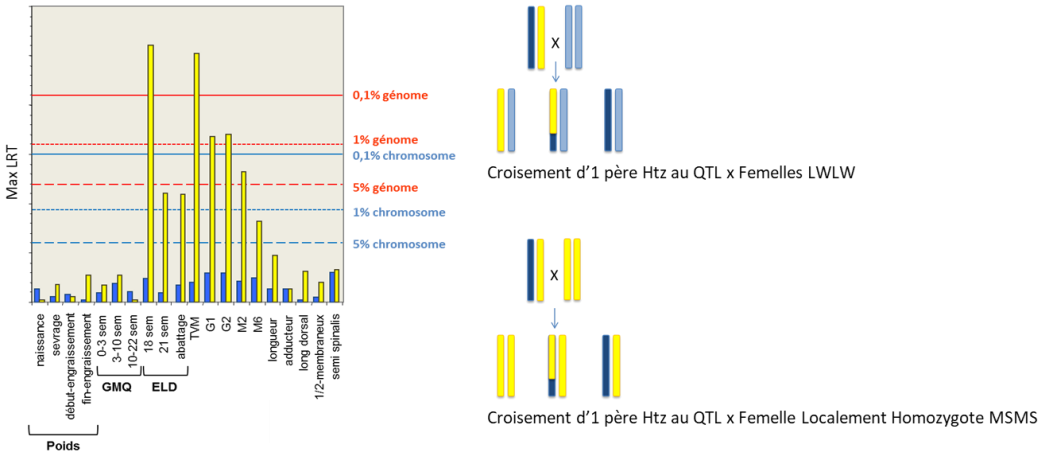


Figure 17 : Validation de la création d'une lignée de femelles localement MsMs.

Les histogrammes en bleu correspondent aux résultats du test statistique de présence d'un QTL dans la région pour le croisement entre un père hétérozygote et des femelles LwLw pour la région du chromosome 1. Les histogrammes en jaune correspondent aux résultats du test statistique pour le croisement entre un père hétérozygote et des femelles MSMS pour la région du chromosome 1 (Juliette Riquet, communication personnelle).

La réalisation de 4 générations de BC et la production de 7 individus recombinants ont permis de réduire de façon importante l'intervalle de localisation du QTL du chromosome 1 de 30 à 8 cM. Deux générations supplémentaires ont permis d'obtenir un intervalle de 2 cM (Riquet *et al.*, 2011b)

Les approches de cartographie fine à l'aide de croisements BC successifs sont longues et coûteuses et chez le porc très peu de travaux de cette nature ont été entrepris. En deçà d'une certaine taille d'intervalle de localisation (quelques cM) la probabilité d'obtenir un individu BC recombinant dans la zone est très faible. Les études familiales deviennent difficiles à réaliser et d'autres approches peuvent être alors envisagées : des études populationnelles via des analyses de déséquilibre de liaison ou des approches de gènes candidats sont alors privilégiées.

2.3 Stratégie de localisation des QTL après 2009

2.3.1 Séquençage du génome de référence du porc

Depuis le début des années 2000, des génomes entiers de l'Homme (2002), la souris (2002), la poule (2004), ont été séquencés.

A partir de 2007, un Consortium International s'est constitué afin d'initier le séquençage d'une première version du génome du porc. Pour cela 2 stratégies complémentaires ont été envisagées.

Dans un premier temps, une sélection de clones Bac, issus de la banque américaine (CHORI-242) générée à partir d'une femelle Duroc et couvrant la totalité du génome avec un minimum de clones (minimum tiling path) a été choisie pour être séquencée, afin d'obtenir la première version du génome porcin.

Dans un second temps, pour compléter l'information issue de ce séquençage, du séquençage aléatoire appelé Whole Genome Shotgun (WGS) a été réalisé. Cette méthode a été popularisée par Craig Venter, lors de la création de son entreprise Celera Genomics pour le séquençage des grands génomes. A l'époque la difficulté majeure (et le challenge) reposait sur le développement d'un algorithme qui était en mesure d'assembler les millions de séquences obtenues.

La stratégie WGS présentait comme principal inconvénient de ne pas pouvoir positionner facilement des fragments comportant des séquences répétées de grande taille, fréquemment rencontrées dans les génomes de mammifères. Cependant, elle s'avère être une technique extrêmement rapide et peu coûteuse par rapport au séquençage de BAC ordonnés.

L'approche combinée de ces 2 techniques a donc permis d'envisager un séquençage peu profond de la banque de BAC (4X). Ces séquences ont alors servi d'échafaudage pour ancrer l'ensemble des lectures obtenues par Whole Genome Shotgun. De cette manière, la première version de la séquence du génome du porc (Draft V9) a été rendue publique en 2009 (Humphray *et al*., 2007). Bien qu'incomplète, cette première version permettait néanmoins de disposer d'une première ébauche de la séquence de référence porcine. Dans un second temps, la qualité du draft a été évaluée voire corrigée en intégrant des informations complémentaires de cartographie : la puce de génotypage illumina 60K constituée de marqueurs SNP de localisation connue sur le draft a été utilisée afin (1) de génotyper des dispositifs familiaux et de construire des cartes génétiques (Tortereau *et al.*, 2012) et (2) de génotyper les panels d'hybrides d'irradiation afin de construire des cartes RH à l'aide de ces marqueurs (Servin *et al*., 2012). Les ordres obtenus entre cartes génétique, d'irradiation et physique (draft de référence) ont été comparés (Servin *et al*., 2012). L'ordre des marqueurs était globalement cohérent, à l'exception de quelques régions qui ont été rectifiées. Cette méthode a permis d'intégrer 72 Mb de séquence génomique sans position initiale dans l'assemblage. L'ensemble de ces nouvelles informations a pu donner lieu à une mise à jour de la séquence de référence version 10.2 (Archibald *et al*., 2010).

2.3.2 Construction d'une puce SNP Haute Densité (60K)

Les SNP (Single Nucleotide Polymorphism) représentent la principale source de variation des génomes (90% du nombre total de polymorphismes humains) (Venter *et al*., 2001). L'acquisition de données de séquençage de plusieurs individus appartenant à différentes races a permis de détecter de nombreux SNP chez le porc. Initialement 550 000 SNP, issus des bases de données et du reséquençage d'animaux de races différentes, ont été testés pour la réalisation de la puce commerciale. Les marqueurs ont été sélectionnés pour leur

informativité multirace, la fréquence de l'allèle minoritaire (MAF pour Minor Allele Frequency supérieure à 5%) et leur qualité d'alignement sur la séquence de référence. Puis 64432 marqueurs ont été sous-sélectionnés pour assurer une couverture homogène du génome (Fan *et al*., 2010) et génotypés chez 158 individus afin de valider cet outil (puce 60K Illumina ®).

Grâce aux puces SNP, il est aujourd'hui possible d'obtenir en quelques jours le génotype de plusieurs individus pour des dizaines de milliers de marqueurs. Ce génotypage sur l'ensemble du génome a de nombreuses applications, comme le diagnostic ou le contrôle d'apparenté des animaux. Il permet également la recherche de loci associés à des caractères. En effet, l'abondance des marqueurs SNP et leur stabilité au cours des générations en font des marqueurs de choix pour les analyses d'évolution des génomes et les études d'association. Suite à la commercialisation des puces de génotypage SNP, les analyses d'association nommées GWAS (Genome Wide Association Study) sont devenues les analyses de prédilection pour la recherche de QTL et ont peu à peu remplacé les analyses de liaison familiale.

2.3.3 Analyse QTL /GWAS chez le porc

Le principe d'une analyse d'association est de comparer la distribution des allèles des SNP avec la distribution des performances d'un caractère chez des individus non-apparentés (Figure 18). Du fait du déséquilibre de liaison entre polymorphismes, les marqueurs proches des variants causaux montrent une corrélation significative (association) avec le phénotype : la ségrégation des SNP à proximité (en déséquilibre de liaison) de la mutation reflète les différences de performances dues aux 3 génotypes possibles à la mutation.

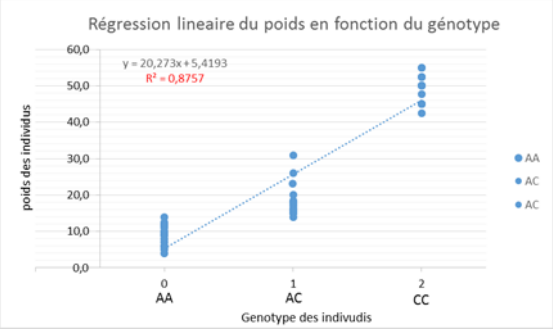


Figure 18 : Principe d'une analyse GWAS.

Recherche d'une association entre un caractère (poids, en ordonnée) et les génotypes des individus (en abscisse) à un SNP d'allèles A/C. Pour ce marqueur, on observe une très forte corrélation (r= 0.94). Dans le cadre d'une analyse GWAS, cette association est évaluée avec chacun des marqueurs répartis sur l'ensemble du génome.

L'utilisation de la puce 60K porcine a ainsi permis de réaliser des analyses d'association tout génome sur des populations commerciales non-apparentées et de définir des intervalles de localisation des QTL plus petits que ceux obtenus par analyses de liaison familiale (Figure 19).

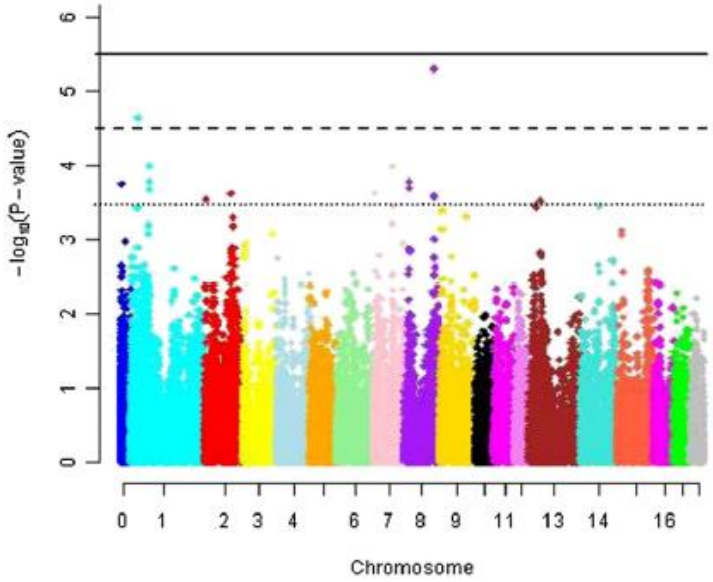


Figure 19 : Représentation graphique d'une analyse d'association tout génome.

Chaque point représente la probabilité qu'un marqueur SNP soit associé au caractère étudié, les 18 chromosomes porcins sont représentés par une couleur différente.

Si avant 2008, un très grand nombre de QTL avaient déjà été mis en évidence chez le porc par des approches de cartographie familiale (466 QTL en 2004), ce nombre a fortement augmenté avec l'utilisation de la puce 60K pour des analyses GWAS. En 2017, 25610 QTL étaient recensés dans la base de données PigQTLdb (version 33 du mois d'août 2017, http://www.animalgenome.org/cgi-bin/QTLdb) (Figure 20). Ces données ont été décrites dans 593 publications et concernaient plus de 646 caractères différents.

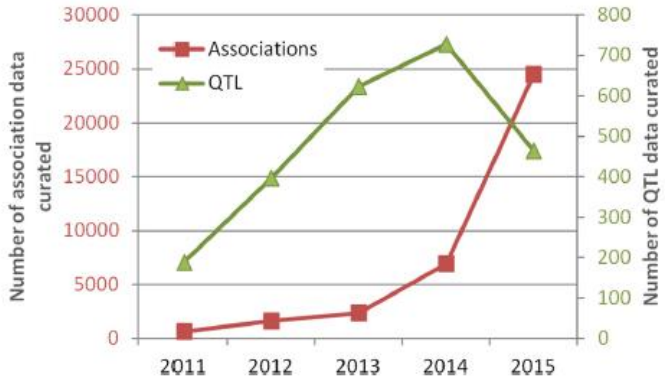


Figure 1. Number of curated QTL and association data per year over the past five years (2015 data are through August).

Figure 20 : Evolution du nombre de QTL détectés entre 2011 et 2015 (Hu et al., 2016).

2.4 Conclusion sur l'évolution de la cartographie de QTL

La cartographie de QTL est basée sur la recherche de marqueurs en DL avec la mutation recherchée et dont la ségrégation présentera ainsi une corrélation forte avec les performances phénotypiques du caractère étudié. Ce DL entre mutation et SNP est dépendant des évènements de recombinaison survenus. La principale différence entre la cartographie familiale et les études d'association repose sur le type de populations étudiées et par

conséquent sur le nombre de recombinaisons exploité dans l'analyse : dans les approches de cartographies familiales, seulement quelques générations de méiose sont mises en jeu (et par conséquent peu de recombinaisons). De ce fait, la résolution de cartographie est limitée et des approches de cartographie fine complémentaires doivent être entreprises pour augmenter le nombre de points de recombinaison.

En revanche, pour les études d'associations, la diversité génétique naturelle et les évènements de recombinaison accumulés au cours d'un très grand nombre de génération peuvent être exploités. Par conséquent le déséquilibre de liaison entre un locus fonctionnel et les marqueurs moléculaires est généralement faible sauf pour les marqueurs localisés à très courte distance de la mutation, ce qui donne une résolution beaucoup plus fine (Zhu *et al.*, 2008) (Figure 21).

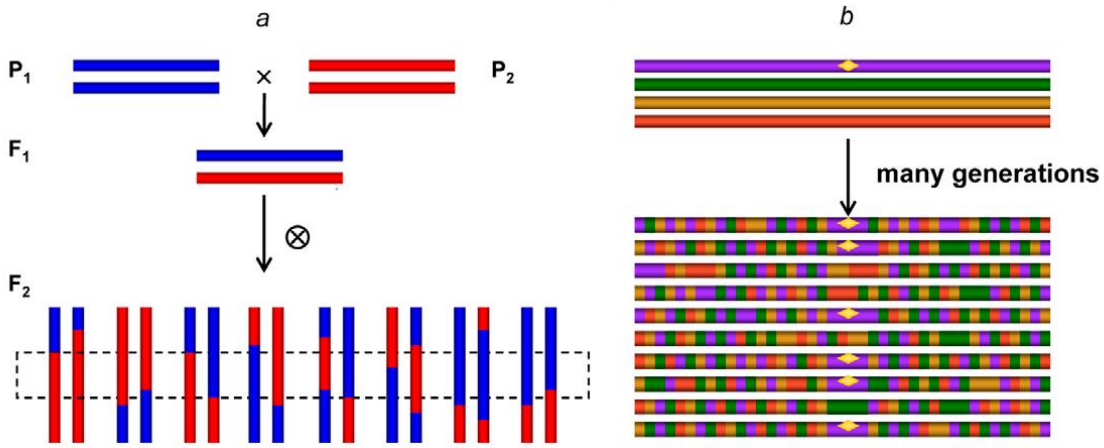


Figure 21 : Comparaison de l'étendue du DL entre les approches de cartographie familiale (à gauche) et les études d'association (à droite). Tiré de Zhu *et al*. (2008).

Ces 2 approches ont donc pour même objectif de déterminer l'intervalle de localisation de gènes gouvernant la variabilité de caractères quantitatifs le plus petit possible. Mais les évolutions technologiques survenues dans les années 2000 ont révolutionné les travaux de cartographie génétique. Via des approches familiales (primo-localisation + cartographie fine), la cartographie d'un QTL dans un intervalle de quelques kb était en général atteinte après plusieurs années de recherche, alors qu'avec les études GWAS, il est maintenant envisageable d'obtenir un intervalle de localisation tout aussi, voire plus, précis en seulement quelques mois.

3 IDENTIFICATIONS DES MUTATIONS CAUSALES

Lorsque l'intervalle génétique devient petit (inférieur à 2 cM), la probabilité d'obtenir un individu BC recombinant dans cette région devient très faible. Les études familiales sont alors difficiles à réaliser et extrêmement coûteuses car elles nécessitent la production d'un grand nombre d'individus. Progressivement des études populationnelles basées sur l'exploitation du déséquilibre de liaison ou des recherches de gènes candidats fonctionnels et positionnels sont alors privilégiées. Parallèlement des études fines sur le caractère sont généralement réalisées afin d'affiner le caractère phénotypique et guider le choix de gènes candidats fonctionnels.

3.1 Phénotypage

Classiquement les mesures phénotypiques réalisées dans le cadre des programmes de sélection ou de détection de QTL sont des mesures qui doivent être facilement réalisables en routine sur un grand nombre d'individus et non invasives pour les animaux. Cependant ces mesures de performances sont généralement peu

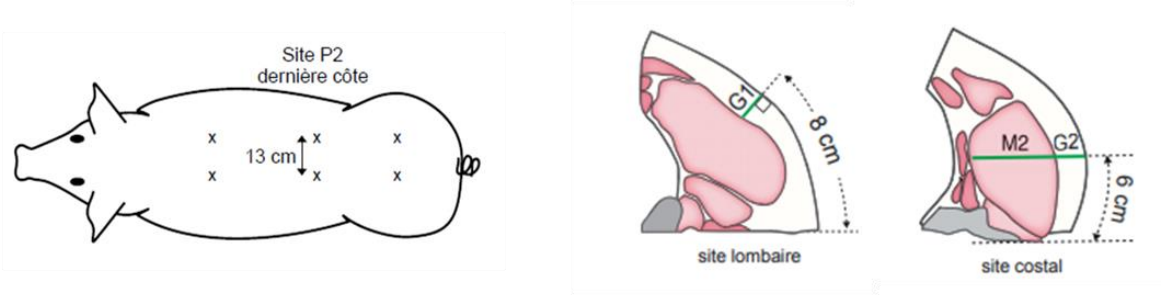
précises car trop globales, comme par exemple le gain moyen quotidien (GMQ) pour l'estimation de la vitesse de croissance ou l'épaisseur de lard dorsal, caractéristique du caractère d'engraissement. Même si ces mesures sont de bons estimateurs des caractères d'intérêt et permettent de réaliser une sélection, elles se révèlent trop imprécises pour étudier les mécanismes moléculaires complexes sous-jacents. Il est donc parfois nécessaire d'affiner la description du caractère par des études physiologiques plus fines.

Dans la suite de cette partie je me limiterai aux mesures phénotypiques du tissu adipeux, caractères affectés par le QTL du chromosome 1, mais la stratégie décrite ci-dessous peut être généralisable à n'importe quel autre caractère étudié.

3.1.1 Mesure du tissu adipeux ou mesure de l'épaisseur de lard dorsal chez le porc

Le phénotypage du caractère d'engraissement est préférentiellement réalisé *in vivo* plusieurs fois au cours de la croissance des animaux. Les mesures sont réalisées à l'aide d'un appareil qui produit des ultrasons (Feat'O mater) ou des échographes, en général au niveau des épaules, du dos et des reins (Figure 22A) à droite et à gauche de la colonne vertébrale.

Des mesures d'engraissement plus fines peuvent également être réalisées à l'abattoir lors de la découpe des pièces en pesant la bardière (tissu adipeux sous-cutané) et la panne (tissu adipeux périnéal) ou en réalisant deux mesures d'épaisseur de lard dorsal suivant deux sites de mesure (Figure 22B) entre les 3 et 4ème dernières vertèbres lombaires, à 8 cm de la ligne médiane dorsale (G1) et entre les 3 et 4ème dernières côtes, à 6 cm de la ligne médiane dorsale (G2).



Mesure *in vivo*
(Feat-O-mater)

Mesure à l'abattoir
(Capteur Gras Maigre (CGM))

Figure 22 : Différents sites de mesure de l'épaisseur de lard dorsal (épaule, dos et rein).

Entre les différentes races de porc, il a été démontré l'existence d'une grande variabilité de ce caractère d'engraissement. Les épaisseurs de lard dorsal de porcs Meishan fluctuent entre 2,5 et 5 cm selon les études, *a contrario* les porcs Piétrains sont extrêmement maigres (ELD Environ 0,7 cm). Les porcs de races conventionnelles comme le Large White ou le Landrace présentent des épaisseurs de lard dorsal intermédiaires (ELD environ 1cm) (Gispert *et al*., 2007).

3.1.2 Description générale du tissu adipeux

Lorsqu'on souhaite mener une approche gènes candidats fonctionnels, il est nécessaire au préalable d'avoir une bonne connaissance du caractère cible. Cela passe donc généralement par des recherches bibliographiques ou des analyses physiologiques complémentaires.

Dans la majorité des espèces de mammifères, deux grands types de tissus adipeux (TA) sont présents: le TA blanc et le TA brun. Le TA brun est principalement impliqué dans la thermorégulation, présent en fin de vie

fœtale et en début de vie post natale chez les mammifères non hibernants et durant toute la vie chez les mammifères hibernants exposés au froid (Cinti, 2005) ; le TA blanc, dont le principal rôle est le stockage énergétique, est présent après la naissance chez les mammifères non exposés au froid (Fève, 2005). Chez le porc, seul le tissu adipeux blanc existe (Trayhurn *et al.*, 1989).

Chez le porc, à la naissance, les dépôts adipeux sont faibles puisqu'ils ne représentent que 1 à 2 % du poids vif (Le Dividich *et al.*, 1991). Par la suite, la croissance est caractérisée par un développement très important du tissu adipeux dont le pourcentage par rapport au poids vif atteint rapidement 12 à 15 % dès l'âge de deux mois et 19 à 23 % vers 105 kg de poids vif (environ 5,5 mois) chez les races à croissance rapide.

Le développement du tissu adipeux chez le porc se déroule en trois phases successives, caractérisées respectivement par une augmentation du nombre d'adipocytes (hyperplasie) entre 1 et 2 mois d'âge (20 kg), une augmentation du nombre et du volume des adipocytes (hyperplasie et hypertrophie) entre 2 et 4,5 mois et une hypertrophie quasi exclusive au-delà, soit à plus de 70 kg de poids vif (Anderson and Kauffman, 1973) (Tableau 3).

Tableau 3 : Développement morphologique du tissu adipeux (TA) chez le porc (1) au cours de sa croissance (d'après Henry 1977).

(1) Porcs mâles castrés Hampshire x Yorkshire

Poids vif (kg)	28	54	83	109
Age (j)	80	117	142	168
Poids du TA dissécable (kg)	3,9	9,7	15,8	25,2
Diamètre des adipocytes dans le TA dorsal (µm)	63,0	78,8	81,5	91,8
Nombre total d'adipocytes dans le TA dorsal ($x10^9$)	25,4	33,2	58,6	64,5

Le tissu adipeux est présent directement sous la peau (sous-cutané) ou au niveau des organes (par exemple le TA périrénal autour des reins), des muscles (TA inter musculaire) ou dans les fibres musculaires (TA intra musculaire). Le tissu adipeux sous-cutané, appelé également bardière chez le porc, représente près de 70% de la masse totale du tissu adipeux, au moment de l'abattage des animaux (environ 110 kg) (Monziols *et al.*, 2006). Environ 20 à 25% du tissu adipeux est présent au niveau intermusculaire. Les dépôts adipeux internes représentent moins de 5% et enfin seulement 1 à 2% du tissu adipeux est localisé au niveau intramusculaire.

Bien que le premier rôle attribué au tissu adipeux blanc soit celui de soutien des organes et d'isolateur thermique, le tissu adipeux est maintenant reconnu comme un organe endocrinien à part entière ayant un rôle clé dans la balance énergétique (rôle essentiel dans le stockage puis la restitution de l'énergie), le métabolisme des lipides, la réponse immunitaire et même la reproduction. Il existe un grand nombre de voies de régulation des principales fonctions métaboliques adipocytaires pour le stockage des lipides (lipogenèse et synthèse de TGs) ou leur dégradation (la lipolyse). Ces différentes voies impliquées sont distinctes et font appel à des substrats spécifiques de chacune d'elles. De très nombreux gènes et facteurs de transcription ont été mis en évidence pour participer à l'adipogenèse et au métabolisme lipidique.

Enfin, les connaissances sur la croissance des tissus suggèrent une priorisation ou une compétition pour la croissance des muscles comparativement à celle du TA. En effet, la croissance du TA augmente quand celle du muscle diminue. Par exemple, la comparaison de génotypes extrêmes bovins culards et de vaches laitières indique que les races bovines diffèrent par leur aptitude de développement des tissus adipeux et des muscles. Les races laitières (Holstein), dont la croissance musculaire s'atténue rapidement avec l'âge ont une proportion de tissu adipeux corporel plus élevée (Gotoh *et al.*, 2009) que les races à viandes (Blanc Bleu Belge, Blonde d'Aquitaine, Limousin, au potentiel de croissance musculaire plus élevé).

3.1.3 Caractérisation plus fine du caractère épaisseur de lard dorsal

La mesure de l'épaisseur de lard dorsal est utilisée pour déterminer le taux d'engraissement des animaux car elle est très facile à réaliser. Cependant cela reste, pour des approches de génétique moléculaire, une mesure phénotypique trop globale. Pour la recherche des gènes impliqués dans la variabilité de ce caractère, il est utile d'avoir recours à d'autres mesures plus fines, afin d'identifier une ou des voies métaboliques impliquées dans le déterminisme du caractère d'intérêt et ainsi orienter le choix des gènes candidats.

Ces mesures complémentaires peuvent être d'origines différentes, comme des mesures fines de cellularité, des dosages biochimiques ou enzymatiques voire des quantifications de niveaux d'expression de l'ensemble des gènes du tissu par des analyses transcriptomiques.

En effet plusieurs travaux associent l'importance de l'épaisseur de lard dorsal chez les animaux de races « grasses » à une hypertrophie des adipocytes. En effet, les races Meishan et basque présentent respectivement des adipocytes plus gros que les animaux de races Pietrain et Large White (Vincent, 2011).

Ce type d'approche a été entreprise dans le cadre de l'étude d'un QTL localisé sur le chromosome 7, influençant l'adiposité des animaux. Dans cette étude, un des résultats les plus surprenants était que les animaux de génotype LwLw au QTL présentaient une épaisseur de lard dorsal plus importante que les animaux de génotype MsMs au QTL. Ce résultat est contraire à ce qui est attendu, les animaux de race Lw étant plus maigres que les porcs de race MS. Pour comprendre ces observations, des mesures métaboliques (métabolismes lipidique, glucidique et oxydatif) et histologiques du tissu adipeux sous cutané dorsal pour les 3 génotypes au QTL ont été réalisées. Ces analyses ont permis de mettre en évidence que les propriétés cellulaires des adipocytes sont modifiées entre les trois groupes d'animaux. En effet, les adipocytes des porcs Lw/Lw sont plus gros que les adipocytes d'individus Ms/Ms, ce qui va bien dans le sens d'une adiposité plus importante (Figure 23). En revanche, aucune différence significative n'a été observée pour toutes les mesures réalisées sur le métabolisme glucido-oxydatif et sur les caractéristiques des fibres musculaires. Tous ces résultats suggèrent que le métabolisme énergétique ne semble pas dépendant de l'allèle Lw ou Ms au QTL, mais les résultats obtenus sur les caractéristiques cellulaires des adipocytes semblent mettre en évidence une implication du QTL d'engraissement dans le processus de croissance des adipocytes.

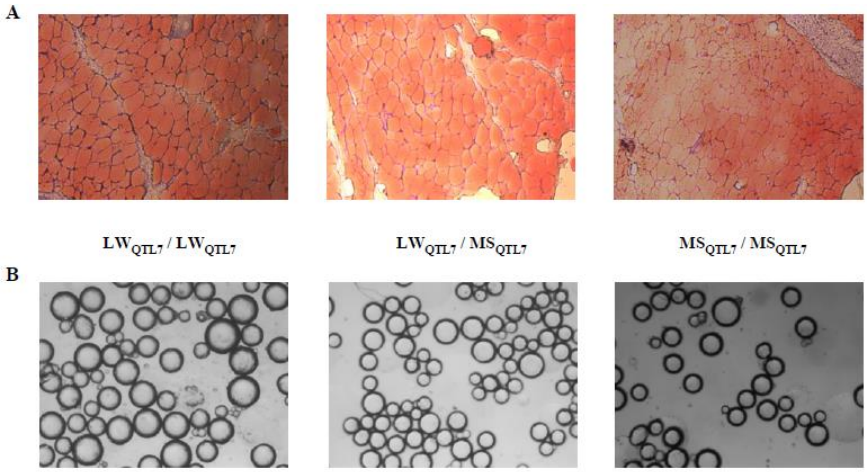


Figure 23: Diamètre des adipocytes du tissu adipeux sous-cutané (TASC) (figure tirée de l'article Demars, 2007).

*Le diamètre adipocytaire a été estimé par deux approches distinctes (à partir de coupes histologiques (**A**) et d'isolement d'adipoyctes (**B**)).*

Bien que ces premiers résultats semblent clairement indiquer un rôle de l'adipogenèse, il n'était pas possible de définir quelle étape (détermination, prolifération ou différenciation) était concernée, ce qui laissait encore un grand nombre de gènes candidats positionnels et fonctionnels impliqués.

Afin de réduire ce nombre de gènes candidats encore trop important, il peut être envisagé de coupler ces études physiologiques fines par des analyses transcriptomiques qui peuvent permettre d'aider l'identification de la voie métabolique impliquée. En effet, l'analyse du niveau d'expression de plusieurs milliers de gènes, à l'aide de puces d'expression (macro ou microarray) sur une centaine d'individus de génotypes différents peut indiquer l'implication d'une voie métabolique. Ce type d'approche se fait sans a priori sur la fonction des gènes, il est cependant évident que cette approche implique de choisir le tissu et le stade de développement à prélever. On pourra conclure que le QTL considéré a des effets sur cette voie métabolique si on observe que le taux d'ARNm de gènes appartenant à une même voie de signalisation semble fluctuer selon le génotype au QTL.

Les données transcriptomiques peuvent alors être utilisées comme une mesure phénotypique « fine » spécifique du QTL d'intérêt et plus discriminante que certaines mesures réalisées au cours des protocoles de testage.

Il existe également d'autres solutions pour diminuer le nombre de gènes candidats fonctionnels avant de pouvoir *in fine* identifier dans leurs séquences une ou plusieurs mutations associées avec le génotype au QTL.

3.2 Identification des mutations candidates

3.2.1 Approches gènes candidats

Des approches génomiques comparées avec l'homme ou la souris sont généralement utilisées pour identifier l'ensemble des gènes présents dans la région d'intérêt. En effet, après avoir clairement caractérisé la région par des approches de cartographie physique (carte d'irradiation) et par cartographie comparée, il est alors possible de connaitre la liste exhaustive de tous les gènes présents dans cet intervalle. Cependant il est important d'avoir une bonne annotation de l'espèce de référence et de l'espèce étudiée.

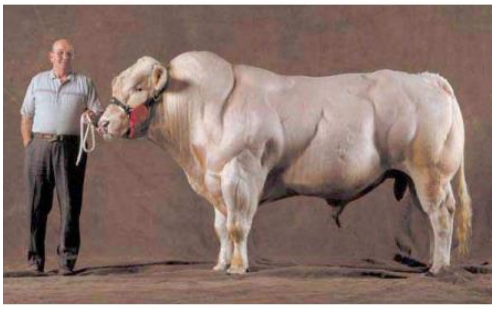
Là encore le nombre de gènes peut rester très important pour les étudier tous individuellement. C'est pourquoi il est nécessaire d'en cibler un ou plusieurs présentant un intérêt particulier sur la base de leur annotation fonctionnelle ou en recherchant leur rôle dans l'expression de phénotypes similaires.

Ce type d'approche a conduit avec succès à l'identification du gène responsable du caractère culard chez le bovin, qui correspond à une augmentation généralisée de la masse des muscles squelettiques (Figure 24). Via une approche classique par analyse de liaison génétique le gène a été localisé sur l'extrémité subcentromérique du chromosome 2 bovin à deux cM du marqueur microsatellite le plus proche (Charlier *et al*., 1995).

Des études indépendantes de génétique inverse chez la souris ont mis en évidence, dans un même temps, un gène qui présentait le même phénotype que les animaux culards. En effet, une augmentation spectaculaire et généralisée de la masse des muscles a été observée chez des souris KO pour le gène GDF8 (McPherron *et al*., 1997), membre de la superfamille des facteurs de croissance et de différenciation de type TGFβ.

Après avoir démontré chez les bovins, à l'aide de la cartographie par hybrides d'irradiation, que ce gène candidat se trouvait bien dans la région souris orthologue à la région bovine définie précédemment, le séquençage de ce gène a mis en évidence une délétion de 11 pb dans la séquence codante de GDF8, avec pour effet un décalage de la phase de lecture et l'apparition d'un codon STOP prématuré qui supprime la partie biologiquement active de la protéine (Grobet *et al*., 1997).

24A



24B

24C

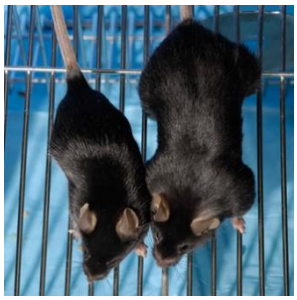


Figure 24 : Différentes espèces porteuses d'une mutation naturelle du gène de la myostatine.

A : Culard charolais (Sélection sur mutation mh).

B : Phénotype « Bully » chez le chien de race Whippet porteur de la mutation mh/mh.

C : Souris témoin et souris KO pour le gène de la myostatine.

3.2.2 Approches IBD

Lorsqu'aucun gène candidat fonctionnel ne semble se dégager, une seconde approche de génétique consiste à tirer parti de l'histoire évolutive des populations. Cette stratégie repose sur le même principe que la cartographie fine par BC mais dans ce cas l'ensemble des évènements de recombinaisons qui ont eu lieu au cours des générations sont exploités, sans avoir besoin de produire des dispositifs particuliers.

En effet, dans ce type d'approche, l'hypothèse est qu'au sein d'un même groupe génétique (comme une race ou une lignée), une mutation est apparue une fois chez un seul individu puis a diffusé au cours des générations. Les individus porteurs d'un même allèle au QTL auront donc tous hérité cet allèle d'un ancêtre commun. Lorsque le chromosome muté sera transmis dans la population au fil des générations successives, les phénomènes de recombinaison vont réduire progressivement la longueur du segment chromosomique. A la nième génération, la recherche d'une portion chromosomique identique chez des animaux présentant le même phénotype permet de définir la région où rechercher le gène impliqué (Figure 25). Plus le nombre de générations séparant les différents animaux issus d'un même fondateur est important, plus le segment chromosomique identique par descendance (IBD) contenant le gène est petit (Bidanel, 2008).

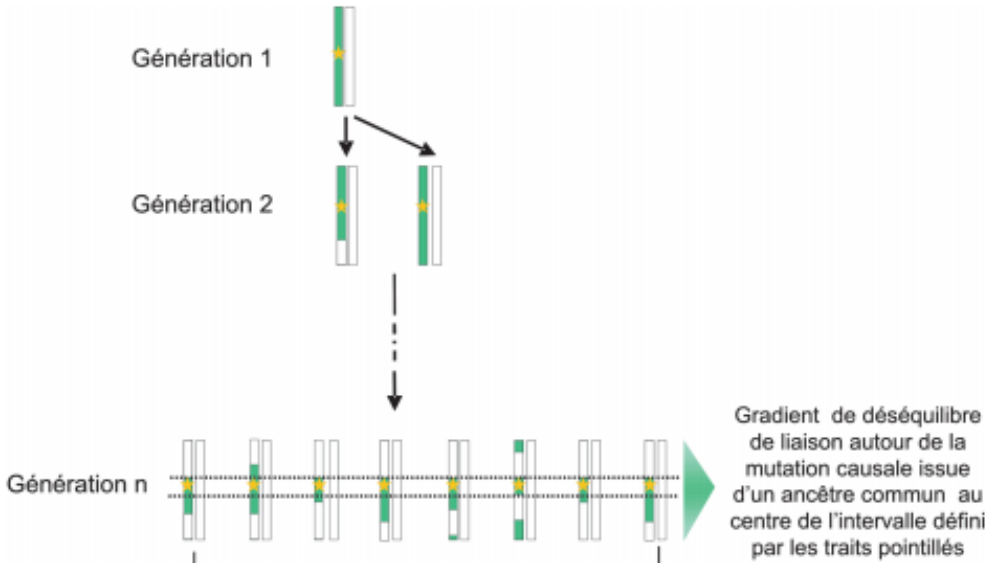


Figure 25 : Cartographie fine de QTL par la recherche de segments identiques par descendances (IBD).

Cependant, ces analyses intra-race ne permettent de réduire l'intervalle de localisation que lorsque les mutations sont très anciennes et qu'un nombre important de générations aura permis de réduire, via les évènements de recombinaison, l'intervalle de localisation. Une autre possibilité complémentaire à cette approche est donc d'utiliser des populations ou des races différentes présentant le même effet au QTL, et de faire l'hypothèse qu'elles sont issues d'un même ancêtre commun, car il est quasiment impossible qu'une même mutation soit apparue dans 2 races différentes.

C'est ce type d'approche qui a été appliquée à l'espèce porcine pour réduire l'intervalle de localisation du locus « IGF2 » influençant le taux de muscle. Dans un premier temps, des individus porteurs du QTL ont été testés sur descendance afin de déterminer précisément leur statut au QTL (hétérozygote Qq ou homozygote qq). Dans un second temps, une analyse haplotypique fine des chromosomes a été réalisée afin de caractériser au mieux les haplotypes présents. La comparaison des haplotypes porteurs de l'allèle « Q » a permis de mettre en évidence que l'haplotype muté Large White et Piétrain dans la région du QTL était identique à un haplotype sauvage d'origine Meishan, à l'exception de la mutation. Cet haplotype partagé peut s'expliquer par l'histoire des races : la race Large White est une race synthétique constituée au cours du 18ème-19ème siècle à partir d'animaux de différentes populations dont des animaux originaires d'Asie. La mutation du gène IGF2 serait apparue dans un chromosome d'origine Meishan introgressé (puis sélectionné) dans la lignée Large White (Van Laere *et al*., 2003).

Cette approche peut s'avérer très puissante pour localiser un QTL ; cependant elle dépend fortement de l'histoire évolutive de la mutation recherchée au sein des races.

De plus, la combinaison, au sein d'une même analyse, d'haplotypes issus de races différentes nécessite que le nombre de marqueurs génétiques soit important pour pouvoir identifier un segment chromosomique IBD. Comparer ainsi plusieurs races nécessitent en général d'obtenir une information exhaustive dans la région étudiée et donc de réaliser le séquençage des individus pris en compte dans la recherche d'un segment IBD.

Lorsque l'intervalle de localisation est réduit le séquençage de la région chez des individus porteurs de l'allèle « Q » et « q » permet de répertorier l'ensemble des polymorphismes candidats. Parmi ces variants, un correspond à la mutation recherchée. L'étape ultime est donc de valider ou invalider ces polymorphismes les uns après les autres.

3.3 Analyse et validation de la mutation

3.3.1 Annotation structurelle

Avant de pouvoir prédire si un variant peut avoir des conséquences fonctionnelles, il est nécessaire d'avoir une annotation précise des gènes sur la séquence du génome. En effet, un gène eucaryote présente une structure morcelée, alternant des régions non codantes, les introns, qui peuvent êtres des régions très longues, avec des portions traduites, les exons (Figure 26).

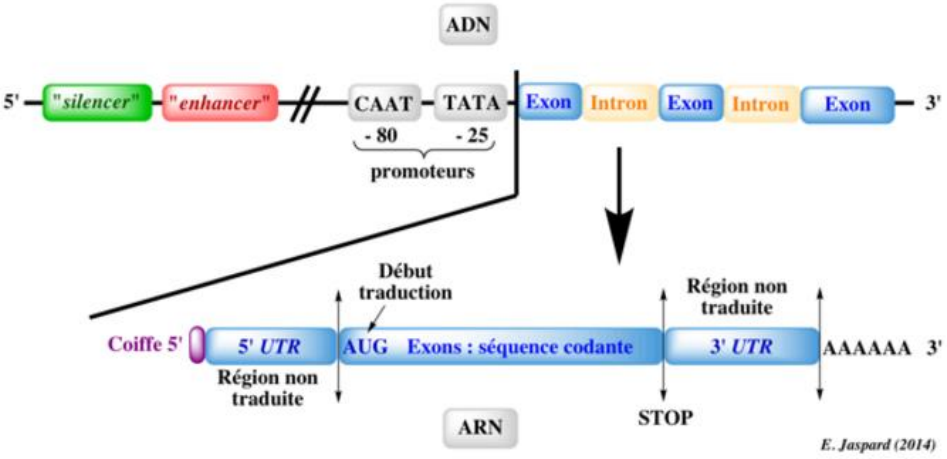


Figure 26 : Structure des gènes chez les organismes eucaryotes.

Lors de la transcription qui se déroule dans le noyau, l'ARN polymérase repère des régions particulières sur la séquence d'ADN (régions promotrices) pour déclencher le processus qui conduit à la fabrication d'un pré-ARN messager contenant les régions non codantes (introns) et les régions codantes (exons). Les introns sont ensuite excisés avant la traduction grâce à un ensemble moléculaire complexe (le spliceosome) qui reconnaît les sites donneurs et accepteurs à la jonction des exons et des introns, et le site de branchement responsable. Enfin, la séquence AATAAA (signal de polyadénylation) pourrait servir de repère de fin des gènes et commander l'ajout d'une queue de plusieurs dizaines de nucléotides A à l'extrémité 3' de l'ARN messager. Cette queue polyA joue un rôle dans la stabilité de l'ARNm, dans son mécanisme de sortie du noyau, et dans la stimulation de l'initiation de la traduction.

L'annotation de la structure des gènes a été principalement définie par des analyses bio-informatiques qui visent à identifier des motifs « consensus » présents dans les régions exoniques, comme par exemple les codons d'initiation ou de terminaison (codon start ou codons stop), les sites donneurs et accepteurs d'épissage, ainsi que les nombreux motifs nécessaires à l'expression des gènes (régions promotrices, facteurs de transcriptions…) localisées dans des régions introniques (Médigue et al., 2002).

Cependant, ces annotations *in silico* sont basées sur des modèles statistiques dont la fiabilité est d'autant plus faible que la taille des exons est petite. Afin d'améliorer la fiabilité des résultats, il est alors conseillé d'avoir recours à plusieurs programmes d'annotation basés sur des algorithmes différents ou à approches expérimentales. En effet, la stratégie la plus efficace pour repérer les régions des séquences codantes reste la comparaison de séquences exprimées (ADNc) à la séquence d'ADN génomique (Foissac *et al.*, 2003). La mise en œuvre de programmes de recherche de similarité, à partir de banques d'ADNc ou d'EST (Expressed Sequence Tags), permet en effet de mettre en évidence les régions de la séquence génomique correspondant aux exons, et de délimiter ainsi les frontières intron-exon.

3.3.2 Annotation fonctionnelle

Une fois que la structure du génome a été clairement définie et que la recherche de l'ensemble des variants a été effectuée, il est alors possible d'analyser ou plus exactement de prédire les effets que peuvent occasionner ces variants dans les protéines, suivant leurs positions dans la séquence.

En effet, des polymorphismes détectés au niveau de certaines régions très particulières peuvent être très intéressants :

▪ **Dans la partie codante**, la mutation peut modifier la séquence de l'ARNm et donc la protéine correspondante. Cependant, du fait de la redondance du code génétique, lorsque la troisième base du triplet est touchée, le variant a de grandes chances de ne pas altérer l'acide aminé (AA) codé ; on parle alors de mutation synonyme. Dans le cas où la mutation provoque un changement d'AA, on parlera de mutation faux-sens. Dans, ce cas, plusieurs cas de figure peuvent être également observés, en fonctions des propriétés physico-chimiques de chaque AA. Enfin, les mutations peuvent aboutir à l'apparition d'un codon stop, provoquant une protéine tronquée non fonctionnelle ou une élimination de l'ARN tronqué.

Plusieurs outils informatiques permettent de prédire les effets fonctionnels des variants, nous en citerons deux. Par exemple SIFT (Sorting Intolerant from Tolerant, http://sift.jci.org) utilise les homologies de séquence entre espèces ainsi que les propriétés physicochimiques des acides aminés pour prédire les impacts des changements ; par exemple, le changement d'un AA hydrophobe par un AA hydrophile en particulier dans le cœur hydrophobe de la protéine devrait avoir des lourdes conséquences sur sa structure. Le second outil très utilisé est PolyPHEN (http://genetics.bwh.harvard.edu/pph2/), ce dernier prend en compte plus particulièrement ces modifications de structure des protéines.

▪ **Dans la partie non codante,** des mutations peuvent également influencer le niveau de transcription de l'ARNm, lorsqu'elles sont localisées dans les régions promotrices ou régulatrices du gène. C'est par exemple le cas de la mutation causale d'IGF2 chez le porc, localisée dans l'intron 3 du gène et qui empêche la fixation d'un inhibiteur de la transcription du gène (Van Laere et al., 2003).

Les mutations peuvent également survenir au niveau des jonctions introns / exons pouvant ainsi entrainer la création de nouveaux sites d'épissage qui seront alors reconnus par la machinerie cellulaire, ou la suppression d'un site d'épissage qui conduira à la production d'une protéine modifiée.

Si des variants sont localisés dans des éléments de régulation tels que les régions promotrices ou dans des sites de fixation de facteurs de transcription, généralement une des premières approches qui est testée pour la validation de ces mutations est l'analyse du niveau d'expression des gènes. En effet, cette méthodologie est relativement simple à mettre en œuvre. Contrairement à l'information génomique qui est identique dans l'ensemble des cellules d'un organisme, le transcriptome varie en fonction du tissu, du type cellulaire, de l'environnement et au cours du temps ; l'élaboration du dispositif expérimental (tissu et condition) reste donc complexe à déterminer.

Cependant la validation fonctionnelle considérée comme la plus probante est de reproduire le phénotype au niveau cellulaire ou chez une espèce modèle, en remplaçant le gène normal par un gène altéré ou non fonctionnel (porteur de la mutation candidate). Il est alors possible d'examiner les effets des perturbations de gènes et d'identifier les mécanismes moléculaires impliqués pour le phénotype d'intérêt. Bien que ces outils permettent d'établir plus facilement des relations de causalité, ces modèles n'abordent pas la question des interactions entre multiples variants génétiques, un phénomène qui est à la base de caractères complexes chez les mammifères.

Par conséquent, si la démarche de cartographie fine qui consiste à localiser, réduire l'intervalle de localisation et enfin inventorier les variants causaux, est une démarche classique et qui peut être généralisée à différents protocoles d'identification de gènes majeurs ou QTL, à l'opposé les étapes suivantes de validations fonctionnelles ne peuvent pas être définies à l'avance. En effet elles dépendent du type de mutation. Les résultats issus de chaque étape influencent la ou les stratégies à mettre en place pour progresser dans la caractérisation de la mutation d'un QTL.

Pour conclure, l'identification de la mutation causale responsable du déterminisme d'un caractère d'intérêt est longue et prend généralement plusieurs années. Si plusieurs mutations associées à des gènes majeurs ont été mises en évidence chez différentes espèces (RN chez le porc (Milan et al., 1996), Booroola chez les ovins (Mulsant et al., 2001) et le gène culard chez les bovins (Charlier et al., 1995)) peu de mutations ont été trouvées pour expliquer des QTL.

Le sujet de recherche de mon diplôme EPHE a donc porté sur la fin de la cartographie fine du QTL localisé sur le chromosome 1 du porc influençant la croissance (% Longe) et l'engraissement (Epaisseur de Lard Dorsal, ELD) des animaux, et l'identification et la validation (ou l'invalidation) des mutations candidates de ce QTL.

Pour répondre à cette question, l'objectif était donc de mener plusieurs approches complémentaires : 1) de produire et d'utiliser des données génétiques pour réduire l'intervalle de localisation de ce QTL ; 2) de diminuer la liste des gènes candidats par des approches de génétique fonctionnelle ; 3) d'identifier l'ensemble des variants de la région par le séquençage exhaustif de la zone.

MATERIELS ET METHODES

CHAPITRE II : MATERIELS ET METHODES

1 LES ANIMAUX

1.1 Les animaux du programme PORQTL

Le dispositif PORQTL est issu du croisement de 6 mâles Large White (Lw) avec 6 femelles Meishan (Ms) afin de produire une première génération d'animaux F1, puis une seconde génération d'animaux F2. Afin d'étudier les QTL affectant les caractères d'engraissement un protocole familial complémentaire de type BC a été mis en place sur l'unité expérimentale GENESI du Magneraud (17). Deux pères F1 PORQTL (F9110001 et F9110012) ont été croisés avec des femelles Lw afin de produire des animaux BC. Pour le QTL du chromosome 1, 3 verrats BC2, issus à l'origine du même père F1 (F9110001), ont été testés sur descendance (Figure 27). A partir de la génération BC4, les mâles recombinants ont été croisés avec des femelles localement congéniques Ms/Ms, comme cela a été mentionné dans la partie données bibliographiques (2.2.3.3 : Création d'individus BC localement congéniques).

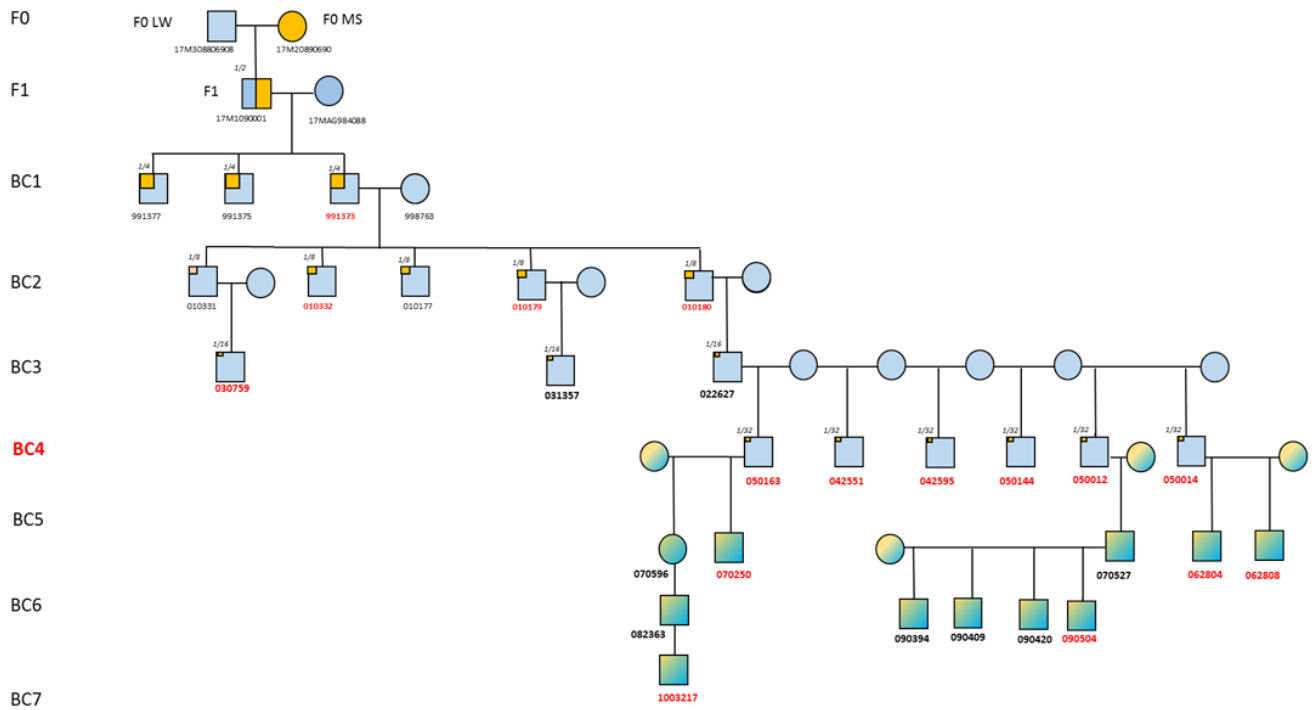


Figure 27 : Pedigree du dispositif PORQTL.

La couleur bleue symbolise des individus mâles ou femelles de race Lw alors que la couleur jaune représente les animaux de la race Meishan. Ce code couleur sera conservé tout au long de ce mémoire. La fraction en haut à gauche de chaque mâle recombinant correspond à la proportion du génome Ms au sein de ces individus. Les symboles qui présentent un dégradé jaune-vert symbolisent les animaux dont il n'est plus possible d'estimer la proportion du génome Ms. En effet à partir de la génération BC4, les pères recombinants ont été croisés avec des femelles localement congéniques, la proportion du génome Ms au sein de ces femelles et des descendants ne peut donc plus être estimée de façon précise. Les animaux dont le nom est en rouge sont les animaux qui ont été testés sur descendance afin de déterminer leur statut au QTL.

Pour réduire l'intervalle de localisation, 7 générations de BC ont été nécessaires et 16 pères recombinants ont été testés sur descendance. Lors du testage sur descendance plusieurs mesures d'engraissement ont été réalisées sur les animaux vivants ou lors de l'abattage afin d'évaluer leurs performances et de déduire le génotype au QTL du père. Les effectifs mesurés sont résumés dans le Tableau 4. Sur les animaux vivants, quatre mesures d'engraissement ont été enregistrées, l'ELD au niveau du cou, du dos, des reins et la valeur moyenne, sur les animaux âgés de 120 jours et 140 jours. A l'abattoir en plus des 4 mesures citées ci-dessus, deux mesures d'engraissement plus fines au niveau de deux sites particuliers, entre les 3 et 4ème dernières vertèbres lombaires (G1) et entre les 3 et 4ème dernières côtes (G2) ont été ajoutées (Figure 22). Les mesures à l'abattoir sont plus compliquées à réaliser, par conséquent les mesures n'ont pas été réalisées sur la totalité des animaux et à partir des BC5, ces mesures n'ont plus été effectuées.

Tableau 4 : Nombre de descendants testés par père.

Père	Génération	Nombre de descendants testés *in vivo*	Nombre de descendants testés à l'abattage
18GAL010179	BC2	148	63
18GAL010332	BC2	144	68
18GAL010180	BC2	79	32
18GAL031357	BC3	134	111
18GAL041462	BC4	152	121
18GAL042551	BC4	101	87
18GAL042595	BC4	94	78
18GAL050163	BC4	100	95
18GAL050144	BC4	61	57
18GAL050012	BC4	177	110
18GAL050014	BC4	139	77
18GAL062804	BC5	119	
18GAL062808	BC5	143	
18GAL070250	BC5	112	
FR18GAL200900504	BC6	82	
FR17MAG2010003217	BC7	72	

1.2 Les Animaux du programme BIOMARK

Des animaux complémentaires issus du programme de recherche BioMark ont été utilisés dans le cadre de ce travail. Ce programme ANR était destiné à évaluer si les QTL identifiés à partir du dispositif PORQTL étaient en ségrégation dans d'autres races et lignées françaises. Dans le cadre de ce protocole 52 verrats de races différentes ont été testés sur descendance, selon le même protocole de phénotypage que celui utilisé pour les verrats BC (Tableau 5).

Tableau 5 : Effectifs des 52 familles de pères testés dans le cadre du programme Biomark

Origine	Nombre de pères	Lignées constituants les familles	Nombre moyen de descendants par père
INRA	8	BC : (Lw x Ms) x Lw	191
INRA	16	BC : (Lw x Pietrain) x Lw	122
INRA	4	F2 : Lw x Pietrain	73
INRA	4	F2 : Lw x Duroc	114
Nucleus	3	Lw	91
ADN	2	Lw	91
Gene+	1	Pietrain	89
Nucleus	2	Pietrain	63
PenArlan	1	Pietrain	108
ADN	1	Landrace	89
Nucleus	1	Landrace	120
Gene+	2	Landrace	96
Nucleus	1	Duroc	110
ADN	1	Duroc	72
Gene+	1	Duroc	56
PenArlan	1	Redon	97
PenArlan	1	Solbec	107
PenArlan	1	Neckar	92
Gene+	1	Taizumu	129
	52		**1910**

2 LES ECHANTILLONS

2.1 Préparation des Echantillons ADN

2.1.1 Extraction d'ADN

Les extractions d'ADN sont réalisées à partir de prélèvements sanguins ou à partir d'un morceau de queue. Le protocole utilisé est une adaptation du protocole Montgomery et al., 1990. Pour ces 2 types de prélèvements les étapes sont similaires, mis à part la première étape pour les échantillons sanguins, qui consiste à lyser les globules rouges afin d'obtenir un culot de globules blanc.

A partir de 4ml de sang, on ajoute 6 ml d'une solution de lyse (NH4Cl 150 mM, KCl 10 mM, EDTA 0,1 mM), puis on conserve dans la glace 45 min, jusqu'à la lyse. Un culot de globules blancs est obtenu après centrifugation de 15 min à 3000 g à 4°C, puis 2 ou 3 lavages sont réalisés avec 5 ml d'une solution de lavage (NaCl 140 mM, KCl 0,5 mM, Tris HCl 0,25 mM pH 7,4).

Par la suite, les étapes sont identiques que ce soit à partir d'un culot de globules blancs ou de 3 à 4 « rondelles » de tissu (de la queue) d'un mm d'épaisseur. Les échantillons biologiques sont digérés en présence d'un mélange de 900 µL de TE 10/0.1, 50 µL EDTA 0.5m M pH8, 50 µL SDS 10% et 10 µL de protéinase K à 20 mg/mL (Proteinase K liquide stabiliséee, Eurobio). Ce mélange est agité légèrement et incubé pendant toute la nuit à 37°C. Ensuite, 400 µL de solution saturée NaCl sont ajoutés, la solution est agitée délicatement et centrifugée 20 min à 12000 rpm à 4°C. Le surnageant est ensuite récupéré et transféré dans un tube propre contenant 3 mL d'éthanol 100%. L'ADN est récupéré sur une tige plastique et mis en présence d'1 mL de TE 10/0.1. Afin de solubiliser l'ADN, la solution est placée 1 h à 65°C puis sous agitation douce toute une nuit à 37°C.

2.1.2 Dosage des d'ADN et contrôle qualité

Après extraction, un dosage spectrophotométrique à l'aide du NanoDrop™ 8000 (ThermoScientific) est réalisé sur l'ensemble des échantillons afin d'estimer leurs quantités et d'avoir une idée globale de la qualité grâce au rapport 260/280 (contamination protéique) et 260/320 (contamination en sels).

Cependant pour certaines analyses et notamment pour des amplifications de grands fragments (PCR Long-Range), il est nécessaire d'effectuer un second contrôle afin de valider l'intégrité la molécule d'ADN. Dans ce cas, une migration de l'ADN génomique dans un gel d'agarose 0.8% préparé dans un tampon TAE 1X (Tris Acetate EDTA) est réalisée. L'ADN est visualisé sous UV grâce au Bromure d'Ethidium (BET), agent intercalant de l'ADN.

2.2 Préparation des Echantillons d'ARN

2.2.1 Prélèvements de Tissus

Les prélèvements de tissus sur les animaux de 10 et 30 kg ont été réalisés à l'abattoir de l'unité expérimentale du Magneraud (Charente maritime) alors que les prélèvements sur les animaux de 110 kg ont été réalisés à l'abattoir de l'INRA de Saint Gilles (Ille-et-Vilaine).

Lors des prélèvements les tissus sont disséqués en morceaux de 1 cm^3 environ puis congelés rapidement dans de l'azote liquide. Les échantillons sont ensuite conservés à -80°C jusqu'à l'extraction de l'ARN.

2.2.2 Extraction d'ARN

L'extraction des ARN totaux à partir de tissus est réalisée suivant une adaptation du protocole du kit Macherey-Nagel. La première étape consiste à réaliser l'extraction des ARN à l'aide du réactif TRIZOL® plutôt que celui préconisé dans le kit (Lysis Buffer RA1). En effet, pour certains tissus et notamment le tissu adipeux l'extraction d'ARN peut se révéler difficile à cause de la présence des lipides qui obstruent les membranes des colonnes d'extractions et peuvent donner des rendements très faibles.

L'utilisation du TRIZOL, qui est composé de phénol et d'isothiocyanate de guanidine, permet de préserver l'intégrité des ARN, tout en dissolvant les composants cellulaires.

L'extraction des ARN totaux est réalisée à partir de 80 à 100 mg de tissu en poudre dans lequel est ajouté 1 ml de Trizol et une bille inox. La solution est homogénéisée 2 fois 2 min à 30 Hz à l'aide du Tissue Lyser MM400 (RETSCH). La séparation des ARN se fait ensuite par ajout de 200 µL de Chloroforme et centrifugation à 12 000g pendant 15 min à 4°C. La phase aqueuse qui contient les ARN est ensuite transférée dans un nouveau tube. A partir de cette étape, la purification des ARN et le traitement DNAse sont réalisés à l'aide de colonnes contenant une membrane de silice et suivant les recommandations du Kit Macherey Nagel_Nucleospin® RNA II. L'élution des ARN totaux se fait en déposant 40 µl d'H_2O exempte de RNAse au centre de la colonne, cette étape est répétée deux fois pour optimiser la quantité des ARN obtenue, en utilisant 40 µL de l'éluat pour une seconde élution.

2.2.3 Contrôle Qualité des ARN totaux

Les ARN sont contrôlés d'une part par un dosage spectrophotométrique à l'aide du NanoDrop™ 8000 (ThermoScientific) pour estimer leur quantité et d'autre part, par migration sur gel d'agarose dénaturant, en présence de formaldéhyde et de formamide, pour valider leur qualité.

Quelques échantillons d'ARN, choisis au vu des intensités des bandes 18S et 28S obtenues sur gel d'agarose pour être représentatifs de l'ensemble des ARN, sont dosés sur puces Agilent, méthode de référence pour l'évaluation de la qualité des ARN. En effet, cette méthode permet d'obtenir une valeur de RIN (RNA Integrity Number). Le principe de la puce Agilent est le même que celui d'une migration sur gel d'agarose : les produits sont séparés selon leur poids moléculaire. Le kit employé est l'Agilent RNA 6000 Nano Kit (Agilent Technologies).

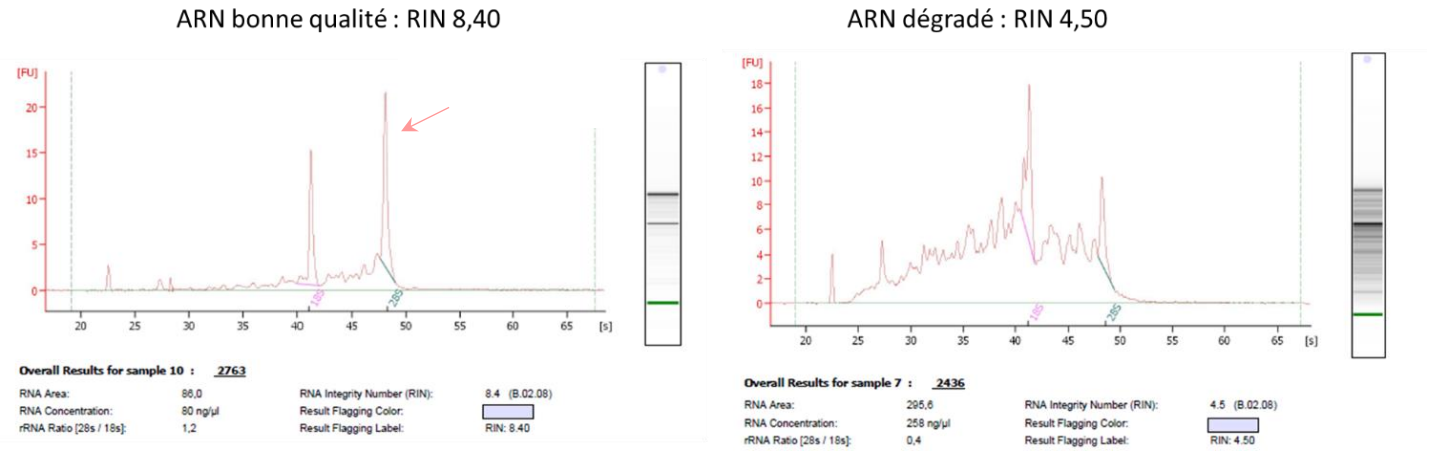


Figure 28: Exemples d'électrophorégrammes présentant des qualités d'ARN (RIN) différents.

Les profils d'ARN présentant un RIN supérieur à 8 présentent majoritairement 2 pics : un fragment 18S et un fragment 28S. Lorsque le RIN diminue, le pic du fragment 28S diminue et de nombreux pics additionnels apparaissent correspondant aux ARN dégradés (Agilent Technologies; RNA Integrity Number (RIN): Standardization of RNA Quality Control).

2.2.4 Synthèse des ADNc

Les ARN totaux ont été transcrits en ADN complémentaires (ADNc), à l'aide de l'enzyme Superscript II First strand synthesis system for RT-PCR (Invitrogen®) en utilisant le protocole fourni par le fournisseur.

Un µg d'ARN totaux sont dénaturés à 65°C pendant 5 min en présence de 1 µl de dNTPs (10 mM) et de 1 µl d'amorces oligodT (15-mer) à 100 µM, puis refroidis immédiatement dans la glace. Le mélange est ensuite incubé 1 h à 42°C en présence de DTT (10 mM), de 200 U de reverse transcriptase Superscript® II (Invitrogen®) ; 40 U de RNAsine et de tampon First Strand Buffer 5X sont ajoutés dans un volume final de 20 µl. La réaction est arrêtée par inactivation de l'enzyme à 70°C pendant 10 min. Les échantillons d'ADNc sont ensuite dilués au 1/5 et conservés à -20°C jusqu'à leur utilisation.

3 TECHNIQUES D'AMPLIFICATION ET DE DETECTION DES ACIDES NUCLEIQUES

3.1 Choix des amorces PCR

La majorité des microsatellites et des SNP ont été développés à partir d'extrémités de BAC ou de séquences disponibles dans les bases de données ENSEMBL (http://www.ensembl.org/index.html) et NCBI (https://www.ncbi.nlm.nih.gov).

Pour l'amplification de fragments compris entre 500 et 1000 pb, les amorces ont été choisies à l'aide du logiciel primer3 (http://bioinfo.ut.ee/primer3-0.4.0/primer3) selon les conditions classiques suivantes : tailles des amorces 18-20 nucléotides, Tm de 58°C, minimiser les appariements entre les amorces.

Pour l'amplification de fragments de 10 kb, les amorces ont été également définies dans primer3, mais pour ces amplifications la taille des primers a été fixée entre 22 à 25 nucléotides et le Tm optimum supérieur ou égal à 65°C.

3.2 Conditions d'amplification par PCR

3.2.1 PCR classique

Classiquement, les réactions d'amplification par PCR pour des fragments compris entre 500 et 1000 pb sont réalisées dans un volume final de 15 µL, composé de Tampon PCR 1X contenant déjà 1,5 mM de MgCl2 (GoTaq Flexi, Promega), 200 µM dNTP, 0.5 µM amorces, 0.5 U d'ADN polymérase (GoTaq, Promega) et 50 ng d'ADN. Les PCR sont réalisées dans un thermocycler GeneAmp PCR System 9700 (Applied Biosystems). Les conditions d'amplification sont les suivantes : 5 min de dénaturation initiale à 94°C puis 32 à 35 cycles, composés de 45 sec de dénaturation à 94°C, 45 sec d'hybridation à la température optimale pour chaque couple d'amorces et 45 sec d'élongation à 72°C, enfin une élongation finale à 72°C et réalisée pendant 20 min.

3.2.2 PCR Long-Range

3.2.2.1 Conditions d'amplification

L'amplification de fragments de grande taille, en moyenne 10 kb, a été réalisée en utilisant une enzyme spécifique et haute-fidélité, l'enzyme polymerase PrimeSTAR® GXL de chez Takara.

La composition du milieu réactionnel et les conditions d'amplifications retenues sont principalement celles recommandées par le fournisseur, à l'exception de la température d'hybridation des amorces.

En effet, nous avons choisi de réaliser une PCR Touch-down : au cours des premiers cycles, la température d'hybridation est diminuée de 1°C tous les 2 cycles. La température initiale est de 68°C, puis elle est descendue progressivement jusqu'à 65°C en 6 cycles, enfin la température d'hybridation est maintenue à 65°C pour les 28 cycles restants. L'avantage de cette hybridation en paliers est qu'au cours des premiers cycles, les produits d'amplification spécifiques prennent quantitativement l'avantage sur d'autres produits qui pourraient apparaitre à une température plus basse.

3.2.2.2 Contrôle Qualité et dosage des produits PCR Long-Range

Dans un premier temps, chaque produit d'amplification est analysé par électrophorèse sur gel à 180 V dans un gel d'agarose 2% préparé dans un tampon TAE 1X (Tris Acetate EDTA) et visualisés par marquage au Bromure d'Ethidium. La taille du fragment est vérifiée grâce au marqueur de taille (1000pb ladder Eurogentec).

Dans un second temps, nous avons souhaité mélanger l'ensemble des produits d'amplification d'un même individu avant de les analyser par du séquençage nouvelle génération (NGS). En effet, nous avons défini 30 couples permettant d'amplifier des fragments de 10 kb en moyenne pour cibler une région de 300 kb.

Afin de mettre la même quantité de chaque amplicon pour obtenir une profondeur homogène tout au long de la région à séquencer, la quantité de chaque produit d'amplification a été estimée sur ce même gel d'agarose à l'aide d'une gamme de 5 échantillons d'ADN commercial de quantité connue (25, 50, 100, 200, 400 ng). En effet, il n'était pas possible d'estimer la quantité par un dosage classique des acides nucléiques (dosage spectrophotométrique ou dosage par intercalant) car certains couples présentaient des bandes parasites (Figure 29). Le principe de ce dosage consiste donc à construire une droite d'étalonnage (régression linaire) à partir de

la gamme d'ADN en relation avec l'intensité des bandes d'intérêt, à l'aide de l'outil Image J pour estimer la concentration de chacune des bandes.

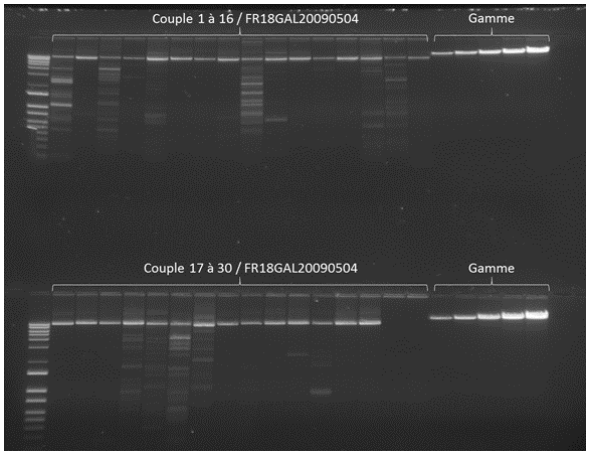


Figure 29 : Résultat de l'amplification des 30 couples couvrant la région de 300 kb pour l'individu FR18GAL20090504.

Le 1ier dépôt à gauche correspond au marqueur de taille de 10 kb. Les 5 derniers puits à droite représentent les dépôts d'échantillons d'ADNg de quantité connue, gamme de 25 ng, 50 ng, 100 ng, 200 ng, 400 ng.

Une fois la concentration déterminée pour chacun des produits d'amplifications, le volume de produits PCR à mettre dans le pool a été estimé pour obtenir une quantité finale de l'ensemble des produits de 2,5 µg, soit environ 85 ng pour chaque amplicon.

Enfin les produits PCR mélangés pour chaque individu ont été purifiés et concentrés à l'aide du Kit CleanUp PCR de chez Macherey Nagel selon les recommandations du fournisseur. Après une purification du volume total du mélange sur colonne, l'élution a été réalisée dans un volume de 45 µl avec le Tampon NE chauffé à 70°C et incubé pendant 5 min avant centrifugation.

3.3 Génotypage des marqueurs de type microsatellite

L'amplification de 9 marqueurs microsatellites a été réalisée selon les conditions PCR classiques décrites précédemment (3.2.1 PCR classique), selon les températures d'hybridation et les concentrations en MgCl2 spécifiques à chaque marqueur (Tableau 6) et à l'aide d'un couple d'amorces dont l'une est marquée à l'aide d'un fluorochrome 6-Fam (bleu), Hex (vert) ou Ned (jaune). Les produits d'amplification ainsi marqués peuvent alors être multiplexés suivant la taille et le fluorochrome utilisé (Tableau 6) avant migration. Pour cela, 5µl de chaque marqueur sont mélangés et on complète avec de l'eau dans un volume final de 100 µL. Ensuite, 2 µL de ce « pool » sont ajoutés à 9,825 µL de formamide et à 0,175 µL de marqueur de taille GS-400HD (Applied Biosystems). Le mélange est dénaturé 5 min à 95°C et la migration est réalisée sur le séquenceur automatique 48 capillaires, ABI PRISM 3730 DNA Analyser (ThermoFisher scientific). Enfin, l'analyse des marqueurs a été réalisée à l'aide du logiciel GeneMapper® v3.0 (Applied Biosystems).

Tableau 6 : Conditions d'amplification et de migration des 9 marqueurs microsatellites

Nom Marqueur	Chr	Pos Genetique (cM)	T° Hybridation	Nb Cycles	Molarité MgCl2 (mM)	Fluorophore	Taille Min (pb)	Taille Max (pb)
MCSE263N22A	1	127,4	58	35	1,5	Ned	253	304
MCS500D4A	1	129,8	58	35	3	6-Fam	163	184
MCS455C8A	1	136	55	35	1,5	Ned	144	192
MCST257C18A	1	136,2	58	35	1,5	6-Fam	175	195
MCSE264C2A	1	138,2	58	35	1,5	Hex	335	382
SW1301	1	140,5	55	35	3	Hex	148	177
MCST97M13A	1	143	58	35	1,5	Hex	180	211
SW2512	1	144	55	35	1,5	6-Fam	101	122
MCS840B11A	1	151	58	35	3	Hex	216	225

3.4 PCR quantitative en temps réel

La PCR quantitative ou PCR en temps réel est une version optimisée de la réaction de PCR conventionnelle, qui permet, grâce à l'utilisation de molécules fluorescentes, une visualisation de l'amplification en temps réel. Cette technique permet donc de déterminer le niveau d'expression de chaque gène pour une condition, un stade et dans un tissu donné.

Chaque réaction PCR est réalisée dans un volume final de 5 µL : 1,5 µL d'ADNc dilué au 1/5 et 2,5 µL de SYBR Green Master Mix 2x (Applied Biosystems) et 150 nM de chaque primer. Cependant, à la différence d'une PCR classique la taille des fragments à amplifier doit être comprise entre 100 et 200 pb et le Tm des amorces est défini de manière à n'avoir qu'une seule température d'hybridation, proche de 60°C. La détection de la fluorescence a été réalisée sur l'appareil ABI7900HT Sequence Detection System (Applied Biosystems) à chaque cycle d'amplification. Enfin, la réaction d'amplification se termine par une étape de fusion, pour cela on passe de 62°C à 97°C par paliers de 0,5°C en faisant l'acquisition de l'intensité de fluorescence en continu. L'ensemble de ces valeurs permettent de tracer la courbe de dissociation et de déterminer la spécificité du produit amplifié. En effet, la présence d'un pic unique traduit la présence d'un seul type d'amplicon (Figure 30*B*).

Suivant les conditions décrites ci-dessus, 24 gènes ont été analysés par PCR quantitative (Tableau 7) : 19 gènes cibles compris dans l'intervalle de localisation du QTL et 5 gènes de référence.

Tableau 7 : Liste des gènes utilisés en PCR quantitative.

Les 19 gènes cibles sont indiqués en bleu et les 5 gènes de référence sont en ocre.

Gene		Nom	Homme HSA9 GRCh38.p10	Porc SSC1 Sscrofa10.2	Nb exons Porc	Primer.up	Primer.dn	Taille cDNA	Taille genomique
FNBP1	Cible	formin binding protein 1	129,887,187-130,043,194	303,945,075-304,034,969	18 exons	CCTGCCATAGGGACCTGTAA	GTATGACGTGGGGACGTAGC	168 pb	4198
GPR107	Cible	G protein-coupled receptor 107	130,053,426-130,140,169			CACGCTGGTGTTCTTTGTTC	CCGTTGGTCACCTTCTTGAC	162 pb	6442
NCS1	Cible	frequenin homolog isoform 2	130,172,578-130,237,304	304,186,077-304,192,568	5 exons	GGAAGGCTCCAAGGCAGA	GTGTGGATGCGGAGAAAGG	178 pb	6672
HMCN2	Cible	hemicentin-2.	130,265,882-130,434,123	304,346,787-304,445,413	78 exons	CCTGGACGAGTGTCAGGTG	GTAGCTGCCACGGGTGTT	193 pb	1136
ASS1	Cible	argininosuccinate synthetase 1	130,444,929-130,501,274	304,454,129-304,508,998	15 exons	CCCTCTACAACGAGGAGCTG	GAGAGATGGGCAAAGTGAGG	184 pb	6134
FUBP3	Cible	far upstream element (FUSE) binding protein 3	130,578,965-130,638,352	304,575,334-304,625,560	18 exons	ACAGCAGGTCGCTTTCTACG	TCACAGAAAACAAAGGTTCTTCA	176 pb	1104
PRDM12	Cible	PR domain containing 12	130,664,594-130,682,981	304,646,521-304,662,114	5 exons	AACTCGCACAACACCTTCCT	TTGTCCAGCGTGTGGATG	197 pb	1517
EXOSC2	Cible	exosome component 2	130,693,721-130,704,894	304,671,176-304,679,921	9 exons	ACCCAGAGGGTGATGCTGTA	CTCCTTACCCCTCCTGTTCC	151 pb	664
ABL1	Cible	c-abl oncogene 1, receptor tyrosine kinase	130,713,946-130,887,675	304,685,889-304,829,672	11 exons	AATCAGCCACCTTCACCAAG	GAGAGTGTAACCGGCAGGAG	162 pb	1993
FIBCD1	Cible	fibrinogen C domain containing 1	130,903,676-130,936,376	304,844,674-304,879,017	7 exons	GGCTCCGTGAACTTCTTCC	ACAAACCCACACCAAAGCTC	192 pb	5120
LAMC3	Cible	laminin, gamma 3 precursor	131,009,082-131,094,473	304,979,408-305,098,340	31 exons	AGAGAGGATGCTGGGAAATG	CTGCGTCTGGCTGGAGAG	159 pb	1057
AIF1L	Cible	ionized calcium binding adapter molecule 2	131,096,542-131,123,145			CAGCGACACCATCTCCTACA	AGGCTGGCAATGTCTCTCTC	150 pb	2469
NUP214	Cible	nucleoporin 214kDa	131,125,561-131,234,670	305,131,050-305,156,826	12 exons	CTCGTTCGGAGACCAGAAAC	GTAAAGGCTGGGGCTGAAC	195 pb	13722
FAM78A	Cible	hypothetical protein LOC286336	131,258,076-131,264,715	305,269,556-305,284,180	2 exons	TGCCGTACCAGGGGTAGTT	GGGCAAAGCCAGAGTCTTC	186 pb	14693
PPAPDC3	Cible	phosphatidic acid phosphatase type 2 domain	131,289,682-131,309,262			CAAGCTCATTGGCATCACC	GAGGCTGGGGCTCGACT	196 pb	17671
BAT2L	Cible	HLA-B associated transcript 2-like	131,394,093-131,500,197	305,377,245-305,439,325	31 exons	GCTTCCACTTTGCCGACAG	GTTGAACGTAGTGTCCCTGCAC	150 pb	2209
POMT1	Cible	protein-O-mannosyltransferase 1 isoform b	131,502,918-131,523,806	305,440,958-305,456,203	20 exons	ATCCTGCTGCTCCCTGTG	CCTGTCCCCGTAGGTTAGTG	153 pb	861
UCK1	Cible	uridine-cytidine kinase 1 isoform b	131,523,801-131,531,259	305,455,972-305,462,254	7 exons	TGACGCAGTACACCACCTTC	GTCGCCGTTGAGGATGTC	152 pb	3877
RAPGEF1	Cible	ruanine nucleotide-releasing factor 2	131,576,770-131,740,074	305,502,184-305,633,144	26 exons	ACTACATCGACGGGAAGGTG	TGTTCCTGGGCTTGATTTTC	195 pb	844
HPRT	Reference					TACCTAATCATTATGCCGAGGATTT	AGCCGTTCAGTCCTGTCCAT		
TBP1	Reference					AACAGTTCAGTAGTTATGAGCCAGA	AGATGTTCTCAAACGCTTCG		
B2M	Reference					AAACGGAAAGCCAAATTACC	ATCCACAGCGTTAGGAGTGA		
TOP2B	Reference					AACTGGATGATGCTAATGATGCT	TGGAAAAACTCCGTATCTGTCTC		
ActineB	Reference					CACGCCATCCTGCGTCTGGA	AGCACCGTGTTGGCGTAGAG		

3.4.1 Choix des gènes de référence

La quantification relative nécessite l'utilisation de gènes de référence qui va servir de normalisateur : ce gène est en effet considéré comme stable dans le modèle présenté, idéalement il ne doit pas subir de changement d'expression en fonction des conditions analysées. Ce ou ces gènes vont donc permettre de compenser d'éventuels biais de PCR provenant :

- de la variation dans la quantité et la qualité des échantillons
- du rendement d'extraction différent entre échantillons
- de la variation des pipetages
- de la variation d'efficacité lors de la synthèse de ADNc à partir de l'ARN total.

Pour cela nous avons testés 5 gènes de référence (HPRT, TBP1, B2M, TOP2B, ActineB) sur 48 échantillons d'ADNc. Pour notre étude le choix des gènes de référence a été réalisé par la méthode Genorm (Nygard *et al.*, 2007), les gènes retenus sont TBP1 pour les échantillons extraits à partir de foie et HPRT pour les échantillons extraits à partir de longe ou de tissu adipeux.

3.4.2 Mesure de l'efficacité des gènes

La méthode que nous avons utilisée pour analyser le niveau d'expression relative de chaque gène cible est celle décrite par Pfaffl (Pfaffl, 2001). Elle permet de prendre en compte les données d'efficacité du gène de référence ainsi que l'efficacité du gène cible contrairement à la méthode classique du 2ΔCt qui considère que l'efficacité du gène de référence et du gène cible sont identiques, à savoir une efficacité de 2.

Nous avons donc déterminé l'efficacité des 19 couples de gènes et des 5 gènes de références. Pour cela, nous avons effectué une PCR en temps réel sur une série de dilutions au ½ d'un pool d'ADNc pour chacun des couples d'amorces. Cette série de dilutions au ½ doit donner des courbes d'amplification décalées d'un cycle de PCR (Figure 30A), si tel est le cas, la réaction a alors une efficacité égale à 2 soit 100%. En pratique, les Ct de chaque point de gamme sont placés sur un graphe en échelle logarithmique et la pente de la droite de régression linéaire (coefficient directeur) passant par ces points reflète le facteur multiplicatif à chaque cycle

d'amplification. La valeur de l'efficacité pour chaque gène peut être alors définie par la formule suivante : E = $10^{-1/pente}$, une efficacité de 100% correspond à une pente égale à -3,32.

L'efficacité de PCR est un facteur important pour la reproductibilité des résultats. Plus l'efficacité PCR est grande, plus la PCR sera robuste, en particulier lorsque l'on travaille avec des échantillons présentant un faible nombre de copies.

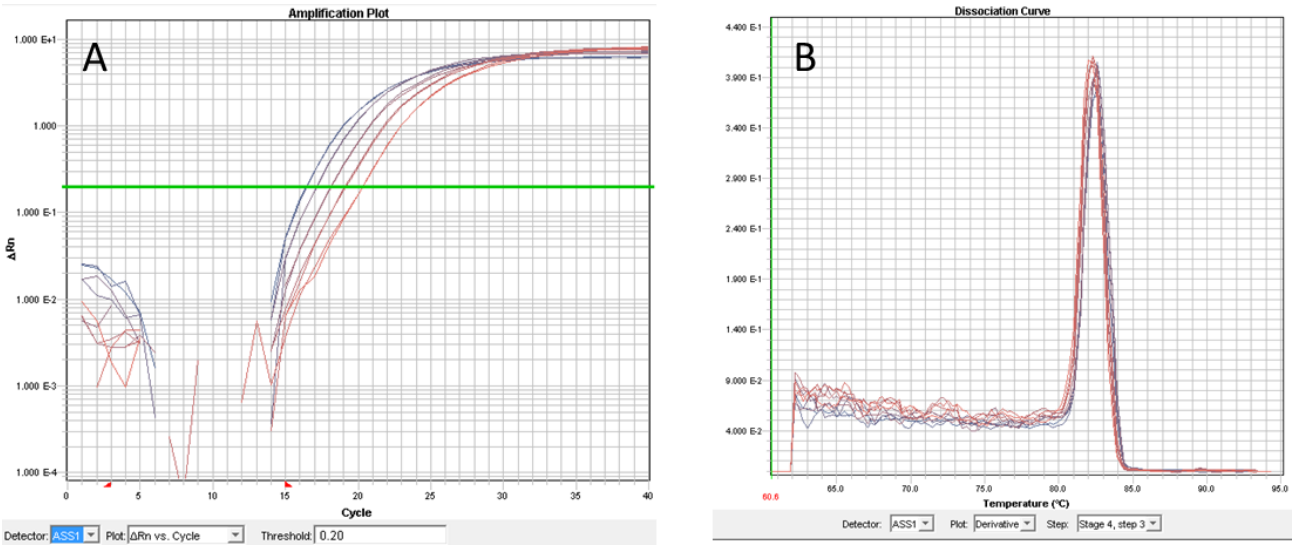


Figure 30 : Profil d'amplification du gène ASS1 sur une gamme d'échantillons d'ADNc poolés et dilués (A) et d'une courbe de dissociation (B).

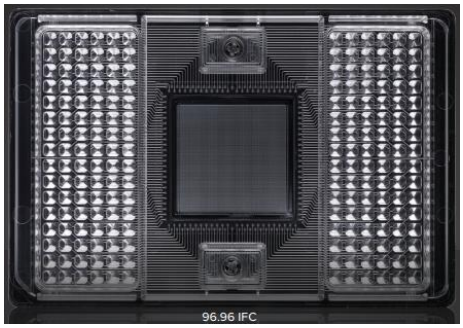
L'efficacité de chaque gène est déterminée à partir de 5 point de gamme réalisé en duplicat à partir d'un pool d'ADNc dilué au demi (1/2, 1/4, 1/8, 1/16, 1/32). Exemple du gène ASS1 pour les échantillons de tissus adipeux (A). B : courbes de dissociation des ADNc double brin en ADN monobrin pour le gène ASS1 et les mêmes échantillons.

3.4.3 PCR Quantitative en Temps réel / Technologie Biomark®

Depuis 8 ans, il existe une technologie de PCR en temps réel Haut-Débit (Biomark®) commercialisée par la société Fluidigm qui permet de réduire les coûts et d'augmenter la fiabilité des résultats. En effet un plus grand nombre d'échantillons et de gènes peuvent être analysés simultanément sur un même support, limitant ainsi les biais d'expérience. Cette technologie permet également de réduire fortement les volumes de réaction, de l'ordre du nanolitre, grâce à un processus miniaturisé basé sur une technologie microfluidique.

Ce système de plaque microfluidique (Dynamic™ Arrays, Figure 31A) est constitué à gauche de 48 ou 96 puits permettant de répartir les différents mix (un mix par gène) et à droite de 48 ou 96 puits permettant de répartir les échantillons d'ADNc qui ont été au préalable pré-amplifiés. En effet, en raison du faible volume final des réactions PCR (7nl), il est recommandé d'augmenter la concentration des gènes cibles dans chaque échantillon. Cette pré-amplification spécifique (STA: Specific Target Amplification) consiste en une PCR multiplex de 10 cycles à 60°C, en utilisant un pool à 0,2 µM de tous les couples d'amorces qui seront utilisés par la suite dans l'expérience. Cette pré-amplification est réalisée à l'aide du kit TaqMan® PreAmp Master Mix. Le mélange de chaque ADNc avec chacun des 48 (96) mix est ensuite réalisé à l'aide de l'IFC (Integrated Fluidic Circuit) dans 2304 ou 9216 micropuits de 7 nanolitres (Figure 31B).

(A)



(B)

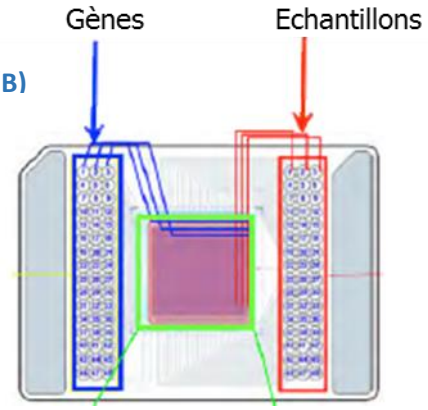


Figure 31 : Plaque Fluidigm Format 96x96 puits et principe de répartition des gènes et des échantillons dans la plaque.

4 LES METHODES DE SEQUENÇAGE

4.1 Séquençage de première génération (méthode de Sanger)

Les régions à séquencer sont obtenues par PCR selon les conditions PCR décrites précédemment, 1 à 12 µl des produits d'amplification (suivant l'intensité des bandes obtenues après contrôle sur agarose) sont ensuite incubés en présence de 0,5 unité d'Exonuclease I (New England Biolabs) et de 0,5 unité de Shrimp Alkaline Phosphatase (Promega) pendant 45 minutes à 37°C. L'exonucléase 1 permet d'éliminer les amorces en excès par digestion de l'ADN simple brin et la SAP (Shrimp Alkaline Phostatase) permet de déphosphoryler le phosphate en 5' des nucléotides libres, empêchant ainsi leur utilisation lors de la réaction de séquence. Ces enzymes sont ensuite inactivées pendant 30 minutes à 80°C.

La réaction de séquençage est ensuite réalisée à l'aide du kit diChloroRhodamine Prism AmpliTaq FS Big Dye Terminator V3.1 kit (ThermoFisher scientific) selon les recommandations du fournisseur, et les produits de la réaction sont déposés sur le séquenceur automatique 48 capillaires, ABI PRISM 3730 DNA Analyser (ThermoFisher scientific).

4.2 Séquençage de seconde génération ou Nouvelle Génération de Séquençage (NGS)

Pour ce type de séquençage, les différentes étapes sont réalisées sur la plateforme Get-Plage de Toulouse par des personnes habilitées, pour cette partie je donnerai donc une description générale du principe.

L'ADN est fragmenté de façon aléatoire par ultra-sons (Covaris). Des fragments d'environ 400 bases sont sélectionnés par migration sur gel d'agarose puis l'ADN est extrait de la fraction du gel prélevé et des adaptateurs (A et B) sont ajoutés aux extrémités de chaque fragment (Figure 32A). La préparation des librairies dure 1 jour et demi. Les molécules d'ADN sont fixées à l'aide des adaptateurs sur une plaque de silice comportant des oligonucléotides complémentaires des séquences A et B. Les fragments sont ensuite amplifiés par PCR localement sur la plaque de silice afin de former des « clusters », puis dénaturés afin de ne conserver qu'un seul brin d'ADN, qui servira de matrice pour le séquençage (Figure 32B).

La réaction de séquençage est réalisée en parallèle pour des millions de fragments par (1) incorporation de nucléotides fluorescents (bloqués en 3'), (2) élimination des nucléotides non utilisés, (3) mesure de la fluorescence émise par le nucléotide incorporé, (4) puis élimination de la molécule bloquante et fluorescente afin de permettre l'incorporation d'un nouveau nucléotide au cycle suivant. Ce cycle est répété une centaine de fois et correspond aux « lectures 1 ». A l'issue du séquençage du brin sens, pour les séquençages en « paired-ends », c'est-à-dire lors desquels on séquence chaque fragment à partir de chacune de ses extrémités, une nouvelle série de cycles de séquençage est réalisée afin d'obtenir le séquençage du brin complémentaire (lectures 2). Le séquençage des brins sens et antisens dure 8 jours au total (Figure 32C).

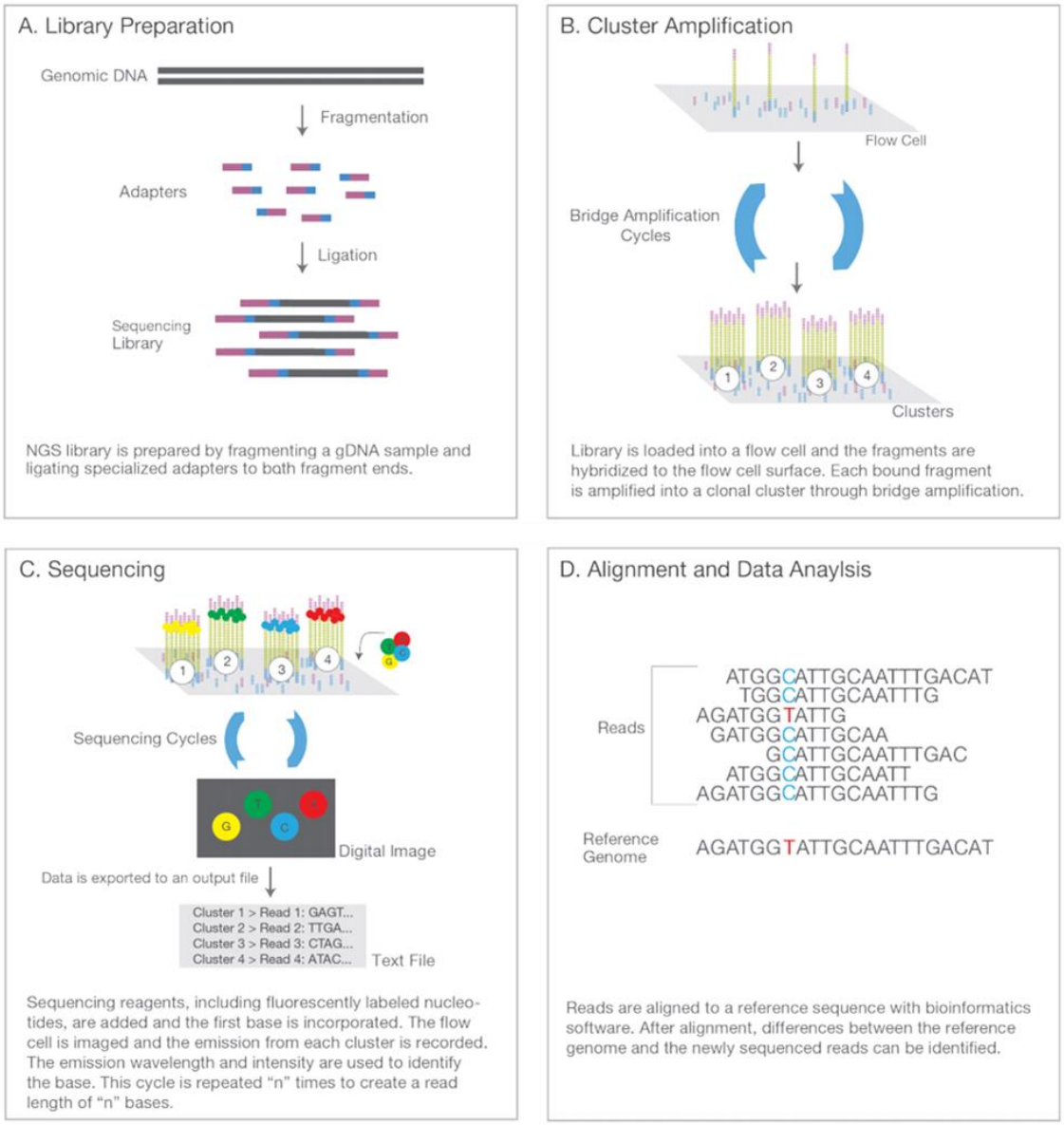


Figure 32 : Les grandes étapes de séquençage Illumina HiSeq3000.

(www.illumina.com/technology/next-generation-sequencing.html)

4 étapes principales : (A) Préparation des librairies, (B) Génération de cluster, (C) Séquencage, et (D) Alignement des lectures et analyse des données.

5 L'ANALYSE DE DONNEES ISSUES DU SEQUENCEUR HISEQ 3000

J'ai réalisé les analyses bio-informatiques avec Patrice Dehais, Ingénieur bio-informaticien de l'équipe de la plateforme SIGENAE (Système d'Information pour l'Analyse des Génomes des Animaux d'Elevage) de Toulouse. Elles ont nécessité dans un premier temps la reconstruction d'un génome de référence et l'utilisation de paramètres spécifiques.

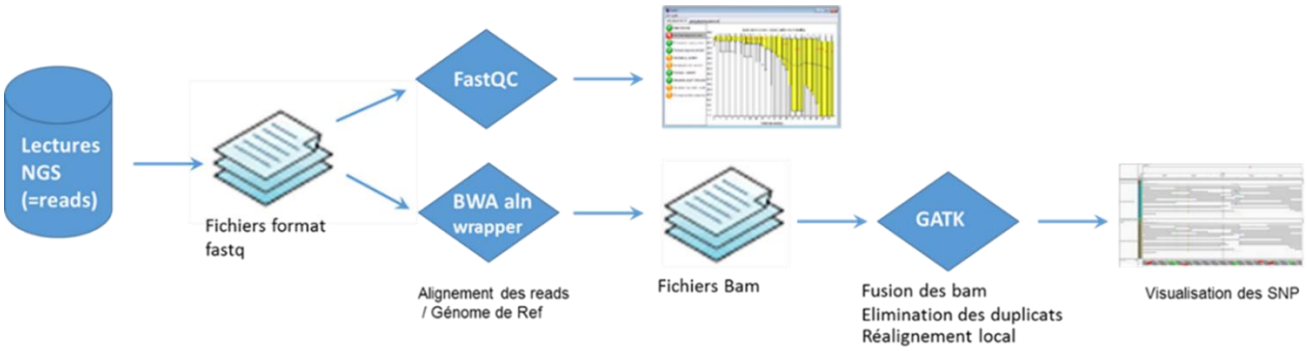


Figure 33 : Pipeline d'analyses et de détection des SNP.

De manière générale, le « workflow » informatique des technologies de séquençage haut-débit appliquées à la détection de variations peut être divisé en 3 grandes étapes (Figure 33) que je présenterai succinctement.

A l'issue du prétraitement des images par le logiciel d'analyses du séquenceur, identification et validation des clusters, des fichiers au format « FastQC » sont générés. Les fichiers FastQC correspondent à la liste des séquences, ou « lectures », appelées également « reads ». Ce format permet de stocker à la fois les séquences nucléiques et les scores de qualité associés. La première étape correspond donc à la validation de la qualité des reads obtenus à partir de ces fichiers fasta.

La seconde étape correspond à l'alignement des séquences sur le génome de référence (BWA aln), Les fichiers générés sont stockés sous forme de fichier SAM (Sequence Alignement/Map) ou BAM (Binary Alignement/Map) (Li *et al*., 2009), ce second format étant la forme compressée du premier format de fichier.

Enfin, la troisième étape est la détection à proprement dit des SNP, ou « variant calling ». Cette étape a été réalisée avec la suite GATK (Genome Analysis Tool Kit) (McKenna *et al*., 2010). Elle consiste à identifier les variations entre l'échantillon testé et la séquence de référence. Le fichier de sortie présente les différents polymorphismes identifiés, selon leur position sur les chromosomes, l'allèle de référence et l'allèle alternatif, ainsi que des valeurs de qualité d'attribution du génotype. La valeur « QUAL » indique la probabilité que le polymorphisme existe, et dépend de la qualité de la base et de la profondeur d'alignement. L'ensemble des résultats (lectures issues des données de séquençage HiSeq, séquence de référence, polymorphismes identifiés, gènes annotés) a été visualisé grâce à l'outil IGV (Integrative Genomics Viewer) (Robinson *et al*., 2011).

6 BASES DE DONNEES ET OUTILS DE BIO-INFORMATIQUE UTILISES

Ces nouvelles méthodes de séquençage se sont accompagnées d'un déluge de données générées en une seule expérimentation comme le mentionne la revue The Economist, en couverture du numéro du 25 Février 2010. De ce fait, les analyses bio-informatiques sont devenues un outil indispensable pour les biologistes. Actuellement, avec les évolutions et le succès des technologies NGS, le nombre de logiciels d'analyse de données de séquençage a augmenté de façon exponentielle. De ce fait devant la multitude de logiciels existants, il est parfois difficile de trouver l'outil approprié. De plus, une des difficultés pour les biologistes réside dans le fait

que l'utilisation de ces logiciels se fait en lignes de commandes et dans des langages de programmation différents et spécifiques (perl, python). Fort heureusement, dans cette multitude de logiciels, les bio-informaticiens commencent à mettre en place des solutions et des outils plus facilement utilisables par des biologistes. Ainsi certaines applications web permettent de simplifier le traitement des données grâce au développement d'interfaces graphiques, de pipelines préétablis. Ainsi The Center for Comparative Genomics and Bioinformatics a développé Galaxy (hhtp : //galaxyproject.org), c'est une « constellation » d'outils pour analyser, manipuler et visualiser des données génomiques haut-débit. Depuis 2012, une instance de cet outil open source a été installée sur les infrastructures locales des serveurs de calcul de la plateforme Genotoul Bioinfo de Toulouse.

De plus, les recherches en génétique moléculaire nécessitent de pouvoir recouper des informations de nature différente et nécessitent donc des interconnections entre les bases. Ces différentes bases sont donc généralement accessibles via des outils d'exploration des génomes, appelés « genome browsers » ou « portails », qui rassemblent des données de tous types issues de plusieurs bases de données indépendantes, et les rendent de ce fait plus facilement accessible. Pour ma part, j'ai principalement utilisé les outils présents dans les 2 « Genome Browser » les plus couramment utilisés, c'est à dire d'une part, le portail du NCBI pour des analyses des séquences nucléiques (les outils Blast, CloneFinder et la base dbSNP) et d'autre part, le portail Ensembl pour les analyses concernant la structure et l'annotation de génomes (Variant Effect Predictor).

Tableau 8 : Principales bases de données et outils de Bio-informatique utilisés.

Nom	Type d'Analyse	Lien du Site
NCBI	Genome Browser	https://www.ncbi.nlm.nih.gov/
blast	Comparaison de séquences	https://blast.ncbi.nlm.nih.gov/Blast.cgi
Clone Finder	Recherche de Clone de BAC	https://www.ncbi.nlm.nih.gov/projects/mapview/mvhome/mvclone.cgi?taxid=9823&build=104.0
dbSNP	Recherche de séquence des variants	https://www.ncbi.nlm.nih.gov/snp
ENSEMBL	Genome Browser	http://www.ensembl.org/index.html
VEP	Prédiction du caractère délétère des mutations Faux-Sens	http://www.ensembl.org/Sus_scrofa/Tools/VEP
GALAXY	Outils d'analyses de séquence haut-débit	http://sigenae-workbench.toulouse.inra.fr/galaxy/
IGV	Outil de visualisation de données de séquençage NGS	http://software.broadinstitute.org/software/igv/
AceView	Base de données d'annotation fonctionnels des transcrits	https://www.ncbi.nlm.nih.gov/ieb/research/acembly/index.html
Genecards	Prédiction et annotation fonctionnelle des gènes humains	http://www.genecards.org/

RESULTATS

CHAPITRE III : RESULTATS

1 ETAT DE L'ART

Cette première partie est destinée à résumer les données et résultats acquis avant le début du stage afin de décrire la situation dans deux domaines : la cartographie génétique et les études fonctionnelles.

1.1 Etat des lieux des données génétiques

Le programme PORQTL initié en 1991 a permis de mettre en évidence un grand nombre de régions, affectant notamment des caractères de croissance et d'engraissement. Les QTL présentant les effets les plus importants ont été localisés sur les chromosomes 1, 2, 4, 7 et X. Cependant à l'issue de ces premières analyses les intervalles de localisation de ces QTL étaient très larges, entre 20 et 40 cM.

La première approche qui a été envisagée pour réduire l'intervalle des QTL est une approche dite backcross ou croisement en retour avec une des deux lignées parentales (en l'occurrence Lw). On obtient alors des animaux qui possèdent un chromosome spécifique d'une race (Lw) et un chromosome recombinant. De ce fait, certains animaux présentent des régions hétérozygotes ou homozygotes dans une partie de l'intervalle de localisation du QTL, et il est alors possible de tester le génotype de l'individu au QTL dans ces régions par un testage sur descendance : homozygote qq ou hétérozygote qQ (l'allèle Q au QTL est associé au chromosome Ms (Meishan) alors que l'allèle q est associé au chromosome Lw (Large White)).

De cette façon, une première série de 4 pères recombinants (3 verrats BC2 et 1 verrat BC3) ont été analysés par testage sur descendance avec des femelles LwLw (q/q) afin de déterminer leur statut au QTL. Cela nous a permis de réduire une première fois l'intervalle de localisation du QTL à 15 cM entre les marqueurs MCS455C8A et MCS840B11A (Figure 34). Cependant, du fait de la dominance partielle de l'allèle Large White et de résultats contradictoires, nous avons été amenés à continuer cette approche en testant les pères recombinants avec des femelles localement congéniques MsMs (cf partie, *2.2.3.3 Création d'individus BC localement congéniques*). La production et le testage de 4 nouveaux pères recombinants ont alors permis de réduire à nouveau de façon significative l'intervalle à une région de 4,3 cM entre les marqueurs MCST257C18A et SW1301 (Figure 34B).

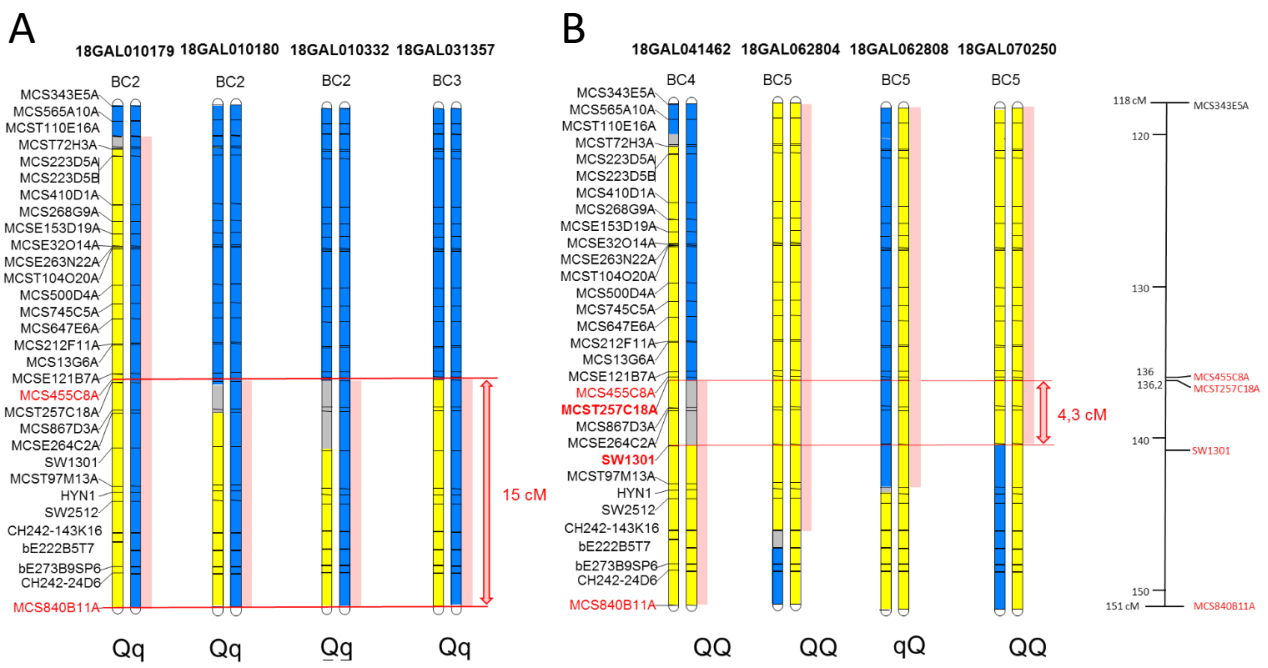


Figure 34 : Résultat du testage sur descendance des pères recombinants dans l'intervalle du QTL.

L'origine raciale des allèles est représentée par un chromosome bleu pour l'origine Lw et un chromosome jaune pour l'allèle Ms. Le statut au QTL des verrats testés est indiqué en dessous des paires de chromosomes (qQ et QQ). L'intervalle de localisation du QTL déduit pour chaque individu testé est alors indiqué par un rectangle rose. Les marqueurs en rouge indiquent les bornes de la région QTL définie par la première série de testage sur descendance et les marqueurs en rouge et gras indiquent les bornes minimales de la région QTL définie après la seconde série de testage sur descendance. La position des marqueurs sur la carte génétique est exprimée en cM.

A l'issue de ces résultats, 2 autres verrats recombinants dans l'intervalle de 4,3 cM étaient encore en cours de testage. Bien que très efficace, les études de cartographie fine par cette approche de croisement en retour sont longues. En effet, il faut compter environ de 2 ans entre la naissance d'un animal Backcross détecté comme recombinant dans la région et le résultat de son testage sur descendance.

C'est pourquoi nous avons fait le choix de mener, en parallèle de ces travaux de génétique, une première approche fonctionnelle.

1.2 Etat des lieux des données transcriptomiques

Bien que ce nouvel intervalle de 4,3 cM semble petit, il comprenait encore une vingtaine de gènes annotés dans la région. Nous avons cependant décidé d'entreprendre une étude systématique du niveau d'expression, sans *a priori* sur la fonction de ces gènes.

Les études transcriptomiques sont des approches assez simples à mettre en œuvre c'est pourquoi elles sont souvent privilégiées lors des premières approches fonctionnelles. Cependant les choix des tissus et des stades sont cruciaux pour espérer pouvoir observer des différences d'expression.

Pour cela nous avons produit spécifiquement des animaux en croisant un père hétérozygote au QTL pour un intervalle de 15 cM (qLw/QMs) avec des femelles hétérozygotes au QTL pour cette même région (qLw/QMs) afin d'obtenir les 3 génotypes possibles au QTL (qLw/qLw, QMs/QMs, qLw/QMs). Pour une trentaine d'animaux de chaque génotype, nous avons choisi de réaliser des prélèvements pour 3 tissus : le tissu adipeux sous cutané, le foie et la longe. La longe peut être un tissu cible pour le caractère de croissance alors que le tissu adipeux et le foie sont 2 organes jouant un rôle majeur dans la lipogénèse mais à des stades différents. De ce fait, nous avons choisi de réaliser ces prélèvements aux 3 stades clés du développement du tissu adipeux chez le porc :

- À 10 kg, au début de la phase d'engraissement correspondant à l'étape d'hyperplasie du tissu adipeux (augmentation du nombre de cellules)
- À 30 kg, pendant la phase d'engraissement, étape d'hyperplasie et d'hypertrophie (augmentation du nombre et du volume des cellules)
- À 110 kg, période de fin d'engraissement durant laquelle l'hypertrophie des cellules perdure, et correspondant au stade durant lequel des mesures d'engraissement sont enregistrées à l'abattoir lors des testages sur descendance.

Ces 3 tissus et ces 3 stades devaient donc nous permettre d'étudier des animaux avec des épaisseurs de lard très différentes.

Dans un premier temps, au vu du nombre d'échantillon à analyser, nous avons décidé de réaliser l'analyse d'expression par PCR quantitative de la totalité des gènes présents dans l'intervalle sur les 2 génotypes extrêmes (LwLw et MsMs) pour le stade 30 kg et les 3 tissus, à l'aide de la technologie Fluidigm.

Lors de l'analyse des 19 gènes par un test de Student, 3 gènes ont montré un différentiel d'expression fortement significatif avec une p-value inférieure à 0,001% pour la longe (FNBP1, GPR107 et HMCN2), 1 seul gène pour le tissu adipeux (GPR107) et 2 gènes pour le foie (AIF1L, POMT1). Le gène GPR107 nous semblait alors particulièrement intéressant car pour les 3 tissus, il présentait un différentiel d'expression, de façon un peu moins significative pour le foie ($p \leq 0.05$).

Tableau 9 : Résultats de la moyenne de l'expression relative pour les 2 génotypes et p-value associée.

Les gènes sont ordonnés en fonction de leur position sur le génome. Les tests les plus significatifs sont représentés en couleur dans le tableau ; rouge = p ≤ 0.001, orange = p ≤ 0.01, rose = p ≤ 0.0 et en gris = p-value non significative.

	Expression relative Longe			Expression relative Gras			Expression relative Foie		
	LWLW	MSMS	p-value	LWLW	MSMS	p-value	LWLW	MSMS	p-value
FNBP1	0,74	1,34	**6,18E-07**	2,32	2,49	0,43	2,41	2,47	0,760
GPR107	1,76	2,71	**6,37E-05**	2,37	3,93	**2,71E-06**	1,74	2,27	**0,004**
NCS1	0,44	0,31	0,03	0,26	0,24	0,71	0,03	0,01	0,18
HMCN2	2,32	1,47	**1,04E-04**	1,10	1,08	0,92	2,30	2,40	0,77
ASS1	0,15	0,12	**0,005**	0,14	0,09	**0,01**	11,44	11,68	0,88
FUBP3	3,05	2,21	**0,01**	2,60	2,44	0,65	2,97	3,05	0,76
PRDM12	5,37	4,55	0,25	2,26	2,28	0,97	1,11	0,72	0,17
EXOSC2	1,83	2,10	0,12	1,71	1,69	0,92	0,99	1,22	**0,03**
ABL1	3,13	2,47	**0,003**	2,33	2,05	0,19	1,12	1,01	0,52
FIBCD1	14,15	12,21	0,18	8,64	7,64	0,43	0,86	0,42	0,25
LAMC3	16,24	12,81	**0,04**	10,86	8,57	0,16	1,13	0,35	0,08
AIF1L	0,11	0,12	0,56	0,44	0,50	0,24	0,20	0,02	**5,54E-06**
NUP214	0,57	0,69	**0,002**	1,39	1,38	0,88	0,68	0,66	0,67
FAM78A	45,11	45,21	0,98	1,36	1,58	0,34	0,73	0,75	0,84
PPAPDC3	29,85	41,02	0,57	3,31	3,54	0,77	0,54	0,27	0,11
BAT2L	1,21	1,12	0,19	2,42	1,93	0,07	0,65	0,50	**0,01**
POMT1	2,00	2,07	0,64	2,69	2,55	0,56	3,12	1,42	**6,50E-04**
UCK1	2,62	2,18	**0,05**	2,54	2,48	0,71	2,05	1,88	0,41
RAPGEF1	3,29	2,62	0,19	1,23	0,86	**0,008**	4,02	4,15	0,92

Ces premiers résultats acquis avant le début de l'EPHE mettaient en avant le gène GPR107, mais se basaient sur une région génétique encore assez grande. Au moment du démarrage du diplôme, nous avons obtenu les résultats de testage des 2 derniers pères recombinants, il nous semblait donc prioritaire de reprendre les analyses génétiques afin de préciser la nouvelle localisation du QTL.

2 RESULTAT DU TESTAGE SUR DESCENDANCE DES DEUX DERNIERS PERES RECOMBINANTS

Comme cela a été décrit précédemment, la stratégie de croisement en retour a déjà démontré son efficacité et a permis d'affiner la cartographie de la région de localisation du QTL de composition de carcasse et d'engraissement.

Au début de la première année de l'EPHE nous avons obtenu les résultats de testage sur descendance des 2 derniers verrats, recombinants dans l'intervalle de localisation de 4,3 cM, correspondant à une distance physique de 1,6 Mb.

Afin de pouvoir déterminer le statut au QTL de ces individus, les verrats FR18GAL20090504 et FR17MAG20103217 ont été croisés respectivement avec 16 et 10 femelles localement MsMs. Ces deux verrats ont permis de produire 96 et 72 descendants, qui ont été génotypés pour un jeu de 9 marqueurs microsatellites de la région (Tableau 6). Pour chacun des descendants des performances d'Epaisseur de Lard Dorsal ont été mesurées par ultrason à 18 et 21 semaines, au niveau du cou, du dos et des reins.

La ségrégation du QTL au sein de ces deux familles a été estimée en testant l'effet de chaque allèle paternel transmis par une régression simple sur les performances de leurs descendants, pré-corrigées pour les co-variables poids, bande de naissance et sexe.

Pour ces deux verrats un effet des allèles transmis très significatif sur les performances d'engraissement a été obtenu. Nous pouvons donc conclure que ces 2 nouveaux animaux sont hétérozygotes pour le QTL d'intérêt localisé sur le chromosome 1.

Le précédent intervalle de localisation du QTL de 1,6 Mb était localisé entre les marqueurs MCST257C18A et SW1301. Le statut « hétérozygote au QTL » de ces 2 nouveaux verrats nous a permis d'exclure les portions chromosomiques homozygote MsMs chez ces 2 animaux et donc de définir une nouvelle borne basse correspondant au marqueur RPI44_276O10 grâce au verrat FR17MAG201003217 et de confirmer la borne haute au niveau du marqueurs MCS257C18A. Ce nouvel intervalle est donc positionné entre les marqueurs MCST257C18A et RPI44_276O10 (Figure 35) et sa taille est estimée à 674 kb.

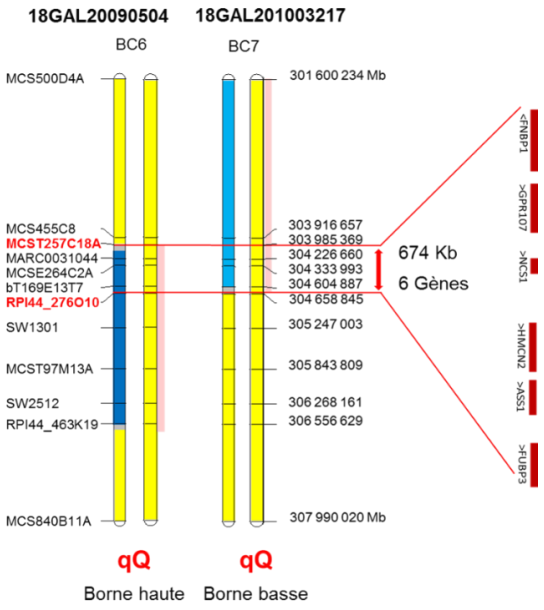


Figure 35 : Intervalle de localisation du QTL défini à l'issue du testage sur descendance des 2 verrats recombinants 18GAL20090504 et 18GAL201003217.

L'origine raciale des allèles est représentée par un chromosome bleu pour l'origine Lw et un chromosome jaune pour l'allèle Ms. Le statut au QTL des verrats testés est indiqué en dessous des paires de chromosomes (qQ et qQ). Les marqueurs en rouge représentent les bornes de l'intervalle de localisation du QTL résultant du testage. Les zones en grisé correspondent aux régions où l'origine raciale des allèles n'a pas pu être déterminée.

Ces résultats sont très importants : ils nous ont permis une fois de plus de réduire de façon très significative l'intervalle de localisation et de définir que seuls 6 gènes étaient encore présents dans cet intervalle : FNBP1, GPR107, NCS1, HMCN2, ASS1 et FUBP3.

3 ANALYSE D'EXPRESSION SUR L'ENSEMBLE DES ANIMAUX DU DISPOSITIF 1

A l'issue de ces nouveaux résultats génétiques, nous avons choisi de compléter les premiers résultats qui avaient été obtenus lors de la première analyse d'expression sur l'ensemble des échantillons du dispositif (3 stades x 3 génotypes x 3 tissus) par PCR quantitative classique (ABI7900HT). En effet, il n'était plus nécessaire de réaliser l'étude d'expression des 20 gènes mais uniquement des 6 gènes encore présents dans l'intervalle.

Après détection du nombre de cycles seuil pour chaque échantillon, nous avons appliqué la méthode PFAFFL pour calculer l'expression relative de chaque gène. L'expression relative a ensuite été normalisée à l'aide de 2 gènes de référence, le gène HPRT pour les échantillons issus de la longe et du tissu adipeux sous cutané (TASC) et le gène TBP1 pour les échantillons de foie.

Les différentiels d'expression ont été ensuite évalués pour chaque stade à l'aide d'un modèle de régression linéaire mixte afin de prendre en compte comme covariables, le poids, l'épaisseur de lard dorsal, le sexe et le numéro de bande des animaux. Ce test nous a permis d'évaluer si le niveau d'expression des 6 gènes étudiés variait en fonction du nombre de copies d'allèle Meishan (0= LwLw ; 1= LwMs ; 2=MsMs).

Tableau 10 : Résultats (p-values et risque beta) du modèle de régression linéaire.

Les tests les plus significatifs sont représentés en couleur dans le tableau ; rouge = p ≤ 0.001, orange = p ≤ 0.01, rose = p ≤0.05 et en noir les p-value non significatives.

Gènes	Stades	Effets	LONGE		GRAS		FOIE	
			beta	pval	beta	pval	beta	pval
FNBP1	10	Geno + Poids + ELD + Sexe	0,1901	0,2022	-0,3766	0,5089	0,1957	0,531
FNBP1	30	Geno + Poids + ELD + Sexe	0,2783	**0,0005**	-0,0021	0,9952	0,1955	0,4699
FNBP1	110	Geno + Poids + ELD + Sexe + Bande	0,1179	0,2421	0,0953	0,8086	-0,6681	0,0348
GPR107	10	Geno + Poids + ELD + Sexe	0,6322	**4,01E-04**	1,8194	**0,0016**	1,5699	**4,78E-07**
GPR107	30	Geno + Poids + ELD + Sexe	1,4168	**1,86E-09**	2,9226	**8,56E-06**	0,8726	**9,55E-05**
GPR107	110	Geno + Poids + ELD + Sexe + Bande	3,336	**1,69E-08**	5,3262	**5,19E-09**	1,6686	**1,53E-04**
NCS1	10	Geno + Poids + ELD + Sexe	0,614	0,4633	-3,747	0,0987	0,0264	0,807
NCS1	30	Geno + Poids + ELD + Sexe	1,4712	0,1614	-2,5116	0,1371	-0,1573	0,2454
NCS1	110	Geno + Poids + ELD + Sexe + Bande	-0,0761	0,9125	6,6867	0,0235	-0,0154	0,8297
HMCN2	10	Geno + Poids + ELD + Sexe	-4,9513	0,367	-11,7606	0,0775	0,0259	0,6448
HMCN2	30	Geno + Poids + ELD + Sexe	-5,8484	0,0485	-5,8194	0,0251	-0,0269	0,7106
HMCN2	110	Geno + Poids + ELD + Sexe + Bande	-0,9516	0,5087	0,9154	0,4947	0,0964	0,0164
ASS1	10	Geno + Poids + ELD + Sexe	-0,0705	0,2067	-0,0779	0,1854	-0,24	0,607
ASS1	30	Geno + Poids + ELD + Sexe	**0,010**	0,8411	-0,1218	0,164	0,0839	0,9162
ASS1	110	Geno + Poids + ELD + Sexe + Bande	0,0259	0,3822	0,0204	0,4446	0,1442	0,8765
FUBP3	10	Geno + Poids + ELD + Sexe	-0,8257	0,0727	0,215	0,2413	0,0116	0,9619
FUBP3	30	Geno + Poids + ELD + Sexe	-1,4253	0,0596	0,2483	0,0536	0,0078	0,941
FUBP3	110	Geno + Poids + ELD + Sexe + Bande	-0,4994	0,383	-0,0274	0,8872	-0,0532	0,7581

Ces analyses nous ont permis de confirmer que le gène GPR107 présentait le différentiel d'expression le plus significatif, pour tous les tissus et tous les stades avec une surexpression de l'allèle Ms (

Figure 36). Pour les autres gènes quelques résultats moins significatifs ont été trouvés, pour FNBP1 dans la longe à 30 kg, NCS1 dans le gras à 110 kg, ASS1 dans la longe à 30 kg et pour HMCN2 dans la longe et le TASC à 30 kg ainsi que dans le Foie à 110 kg (Tableau 10).

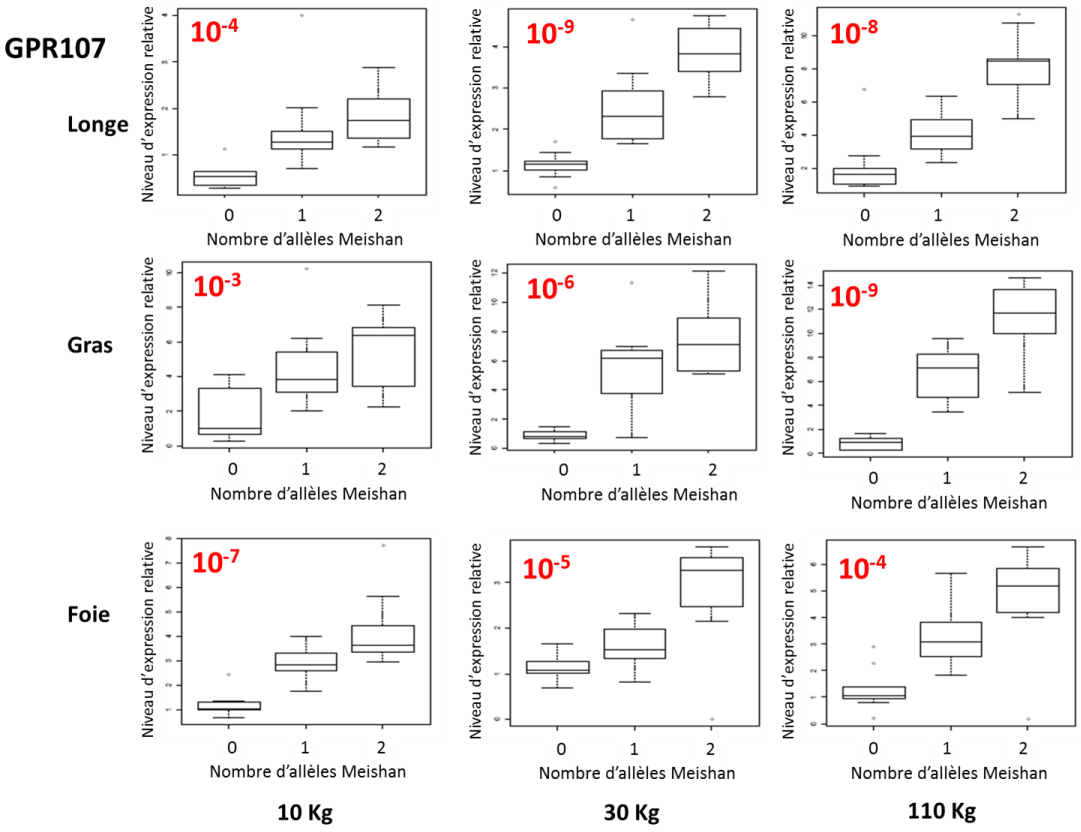


Figure 36 : Boxplot du niveau d'expression relative du gène GPR107.

Box-Plot des niveaux d'expression obtenus pour le gène GPR107 pour chaque tissu et à chaque stade en fonction du nombre de copies d'allèle Ms présent dans la région du QTL. Les p-values de l'effet de l'allèle Ms déterminées par régression linaire sont indiquées en rouge.

Il est cependant important de noter que lors de cette première expérience les animaux qui ont été utilisés pour produire ce dispositif expérimental étaient hétérozygotes pour un intervalle qui s'étendait au-delà de l'intervalle minimum de localisation du QTL défini par les 2 derniers pères recombinants. Il est donc possible d'envisager que la mutation responsable du caractère ne se retrouve plus dans cette nouvelle région.

Afin de s'assurer qu'on observe toujours un différentiel d'expression entre les 2 génotypes dans la région QTL réduite, nous avons donc souhaité mettre en place un second dispositif en utilisant cette fois le père recombinant FR18GAL20090504 qui définit la borne haute minimale, afin d'évaluer si le différentiel d'expression observé pour le gène GPR107 était conservé.

Là encore, le temps nécessaire pour la production des animaux de ce second dispositif étant de plusieurs mois, nous avons pendant cette période, cherché à caractériser les points de recombinaison pour définir précisément l'intervalle minimum de localisation du QTL.

4 RECHERCHE DE L'INTERVALLE MINIMUM DE LOCALISATION DU QTL

Lors du testage sur descendance, la densité du jeu de marqueurs microsatellites utilisé pour le génotypage étant très faible, des zone d'incertitude (région pour laquelle l'origine Lw ou Ms n'est pas caractérisée) subsistent autour des points de recombinaison. Afin de définir le plus précisément possible l'intervalle de localisation du QTL, il était nécessaire de déterminer finement la position des points de recombinaison des 2 derniers verrats testés sur descendance.

Pour déterminer le génotype des deux verrats dans les régions d'incertitude, nous avons choisi d'utiliser des marqueurs de type SNP, plus fréquents dans le génome que les marqueurs microsatellites.

Nous avons tout d'abord retenu les 16 SNP présents sur la puce de génotypage Porcine 60k, entre les 2 marqueurs qui bornaient l'intervalle minimum. Cependant, ces marqueurs n'étaient pas répartis de façon homogène et ils couvraient assez mal les points de recombinaison.

Nous avons alors souhaité développer des marqueurs complémentaires, à partir de la séquence disponible dans la région. Malheureusement, la séquence génomique disponible dans les bases de données (Genome assembly : Sscrofa10.2) pour cette région est composée de 3 contigs séparés par 2 régions de taille inconnue sans séquence disponible.

Nous avons donc choisi de privilégier le développement de nouveaux marqueurs à partir des séquences de BAC localisés dans la région et disponibles dans les bases de données.

4.1 Développement de marqueurs SNP à partir de clones BAC couvrant l'intervalle de localisation du QTL

Le séquençage du génome porcin initié en 2007 était initialement basé sur le séquençage de clones BAC couvrant la totalité du génome. Dans le cadre d'un consortium international, un minimum tiling path de clones BAC a donc été construit. J'ai donc recherché l'ensemble des BAC porcins présents dans la région du QTL à l'aide de l'outil CloneFinder disponible sur le site du NCBI (https://www.ncbi.nlm.nih.gov/projects/mapview/clonefinder/clonefinder.cgi). J'ai ainsi pu récupérer la séquence de 6 BAC (Figure 37), 4 BAC validés et 2 BAC en cours de séquençage (1 BAC sous forme de 2 contigs, 1 BAC comprenant 60 contigs non ordonnés). J'ai privilégié le choix de BAC qui provenaient de la banque de clones qui avait été utilisée pour l'obtention de la séquence de référence, obtenue à partir du séquençage d'un porc de race Duroc. La séquence porcine étant incomplète pour cette région, j'ai décidé de m'aider des données de séquence humaine disponibles afin de sélectionner et d'ordonner les séquences porcines des BAC couvrant l'intervalle du QTL. En effet, de nombreux travaux de cartographie comparée ont montré par le passé que le génome humain pouvait être utilisé comme référence pour des travaux de génomique chez d'autres espèces de mammifères (Goureau *et al*., 1996).

Lors de travaux préalables à l'EPHE, nous avions réalisé une carte comparée de la région du QTL entre ces deux espèces. Une carte d'hybrides d'irradiation construite à l'aide de marqueurs définis à partir des gènes humains localisés sur le chromosome 9 nous a permis de préciser l'homologie entre la région chromosomique humaine 132-134 Mb sur HSA9 et la région porcine 304-306 Mb du chromosome SSC1. J'ai donc comparé avec le programme Blastn les séquences de chaque BAC avec la séquence humaine (GRCh38.p2 reference assembly top-level), pour positionner la séquence des BAC sur la séquence humaine. En effet, le BAC CH242-344M23, composé de 60 contigs, correspondait en partie à une des deux régions absentes de l'assemblage porcin de la région.

Ces comparaisons m'ont ainsi permis de sélectionner 4 BAC couvrant les deux régions de recombinaison (2 pour chaque point de recombinaison, Figure 37). Pour développer de nouveaux marqueurs, 13 couples d'amorces ont été développés dans les régions d'incertitude, espacés en moyenne de 25 kb.

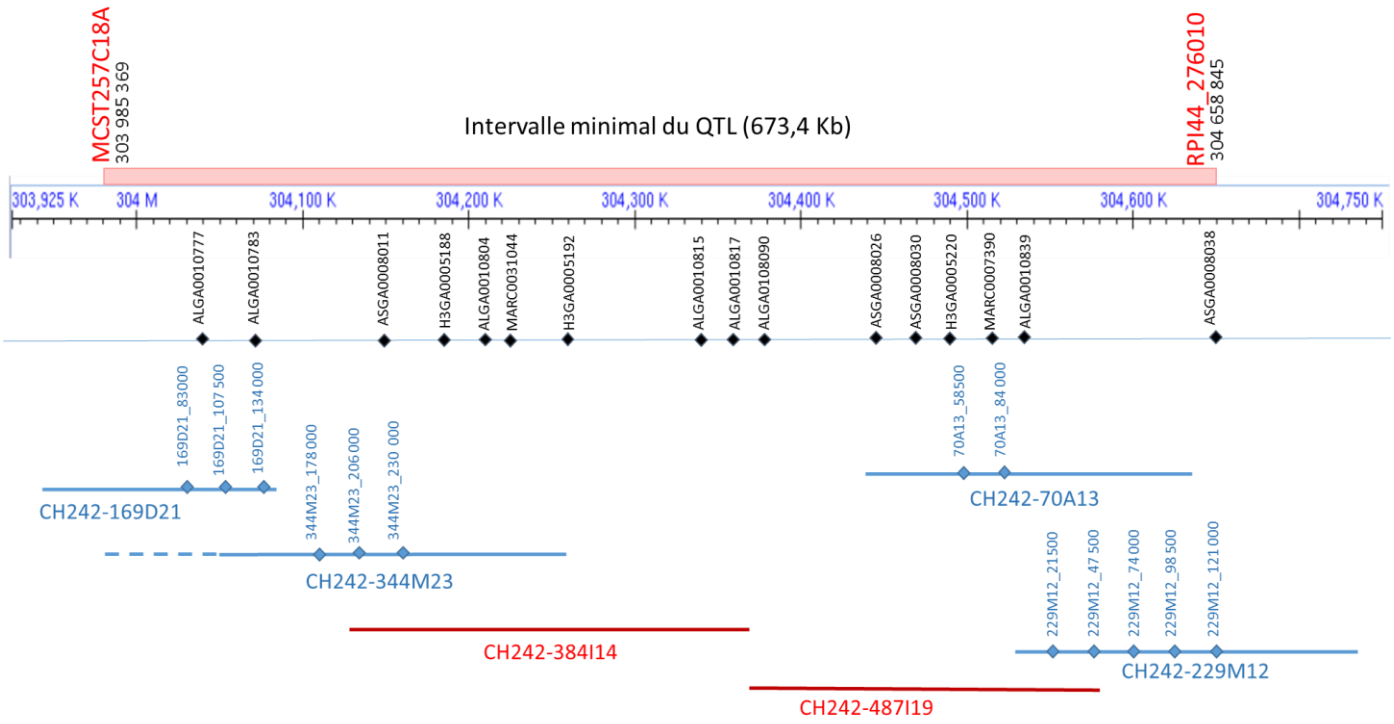


Figure 37 : Origine et localisation des 29 marqueurs SNP utilisés pour densifier la zone QTL.

L'intervalle minimal de localisation du QTL de 673,4 kb est défini par les marqueurs MCST257C18A et RPI44_276O10 indiqués en rouge sur la figure. Les marqueurs en noir correspondent aux marqueurs présents sur la puce porcine 60k alors que les marqueurs en bleu correspondent aux marqueurs qui ont été développés à partir des séquences de BAC. Les BAC en rouge correspondent aux BAC présents dans l'intervalle mais à partir desquels nous n'avons pas défini de nouveau marqueurs.

4.2 Génotypage des marqueurs SNP complémentaires

De cette façon, 29 nouveaux marqueurs ont pu être ajoutés dans l'intervalle MCST257C18A /RP276O10 (Figure 37). Pour ces 29 marqueurs, nous avons réalisé le génotypage des deux individus recombinants et d'un individu MsMs afin de déterminer pour chaque SNP l'allèle Ms.

Le résultat des génotypages de ces 29 nouveaux marqueurs a permis de montrer que l'animal FR18GA20090504 est homozygote Ms/Ms jusqu'au marqueur SNP MARC0031044, alors que l'animal FR17MAG20103217, est homozygote Ms/Ms, à partir du marqueur 229M12_74000 (Figure 38).

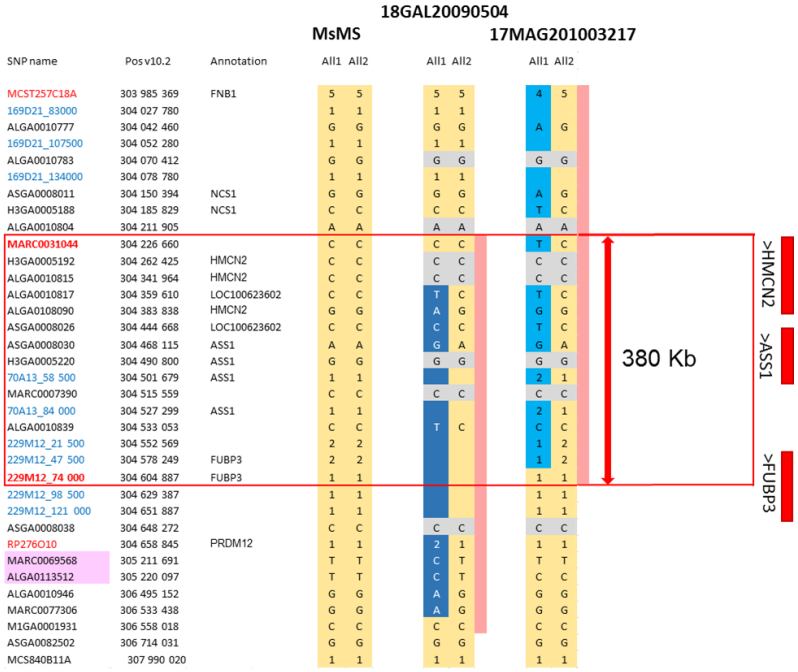


Figure 38 : Précision des points de recombinaison pour les 2 verrats recombinants.

Génotypes obtenus pour les deux verrats recombinants et un témoin MsMs dans la région du QTL pour les 29 marqueurs SNP. Les marqueurs en bleu correspondent aux marqueurs développés à partir de BAC, les marqueurs en noir correspondent aux SNP présents sur la puce 60K dans l'intervalle. Les marqueurs en rouge délimitent l'intervalle de l'ancienne localisation du QTL (MCST257C18A/RPI44_276O10) alors que les marqueurs en rouge et gras définissent les bornes du nouvel intervalle de recombinaison (MARC0031044/229M12_74000).

L'hétérozygotie au QTL des 2 verrats recombinants nous a donc permis de préciser la position des points de recombinaison entre ces 2 nouveaux marqueurs et par conséquent de réduire la région du QTL à une zone de 380 kb, qui ne contient plus que 3 gènes (HMCN2, ASS1 et FUBP3).

Bien que la réduction très significative de la région QTL soit un résultat extrêmement positif, ce résultat a été également très surprenant car il exclut le gène GPR107 qui jusqu'alors, au vu des analyses fonctionnelles, semblait être le meilleur gène candidat.

Nous avons donc poursuivi les analyses transcriptomiques sur les animaux du second dispositif que nous avions mis en place, pour évaluer si le différentiel d'expression de GPR107 obtenu à partir du premier dispositif était conservé alors que la totalité du gène était exclue de l'intervalle de localisation du QTL. Le second objectif était d'analyser si parmi les 3 gènes restants, un différentiel d'expression était observé.

5 ANALYSE D'EXPRESSION SUR L'ENSEMBLE DES ANIMAUX DU DISPOSITIF 2

Afin d'évaluer plus précisément si la mutation recherchée pouvait influencer l'expression des gènes de la région, une analyse complémentaire du niveau de transcription des 3 gènes localisés dans le nouvel intervalle (HMCN2, ASS1 et FUBP3) et du gène GPR107 a été réalisée à partir d'échantillons obtenus sur les animaux du second dispositif expérimental. Pour ce dispositif, le verrat FR18GA20090504 a été croisé à 7 femelles, hétérozygotes LwMs à l'extrémité du chromosome 1. Des descendants de 4 génotypes possibles ont été obtenus (Figure 39), mais seuls les descendants MsMs et RecMs ont été utilisés dans la suite de l'étude, puisqu'ils peuvent être comparés aux animaux du dispositif 1.

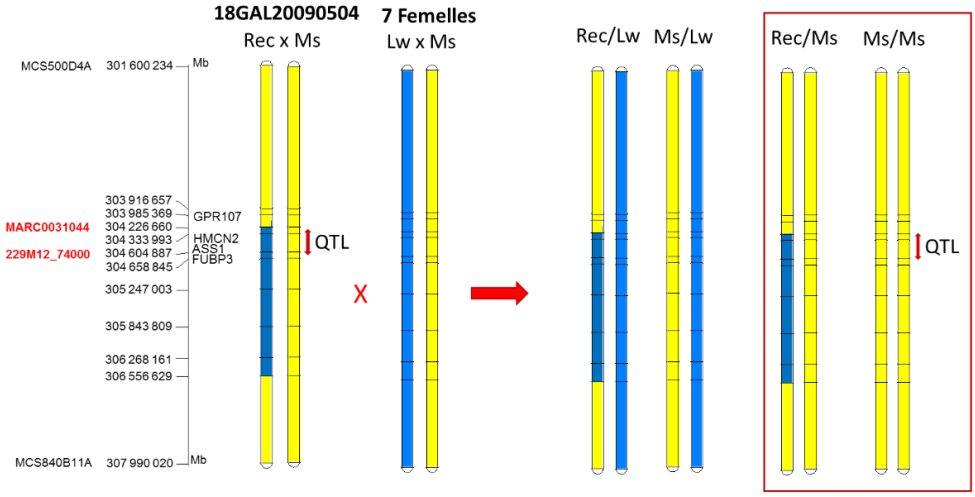


Figure 39 : Second dispositif expérimental pour l'analyse du niveau d'expression des 6 gènes présents dans l'intervalle.

Représentation schématique des génotypes obtenus dans le second dispositif. Le croisement a été réalisé à partir du père recombinant 18GAL20090504 et de 7 femelles LwMs. Ce croisement a permis d'obtenir 4 classes de descendants. Chaque classe est représentée par une paire de chromosome d'un individu type ; l'origine raciale des allèles est indiquée par une couleur (Lw : bleu et Ms : jaune).

En effet l'objectif majeur était d'évaluer si le différentiel d'expression du gène GPR107 observé à partir des animaux du premier dispositif transcriptome était (1) corrélé à la ségrégation de la mutation recherchée ou (2) résultait d'un polymorphisme en amont ou dans le gène (régions désormais exclues via le testage sur descendance du père FR18GA20090504). Sous la première hypothèse, un différentiel devait être observé entre des descendants MsMs et RecMs ; selon la seconde hypothèse les niveaux d'expression devaient être équivalents chez les deux lots de descendants, l'ensemble des animaux étant MsMs dans la région promotrice et les régions codantes du gène (Figure 40).

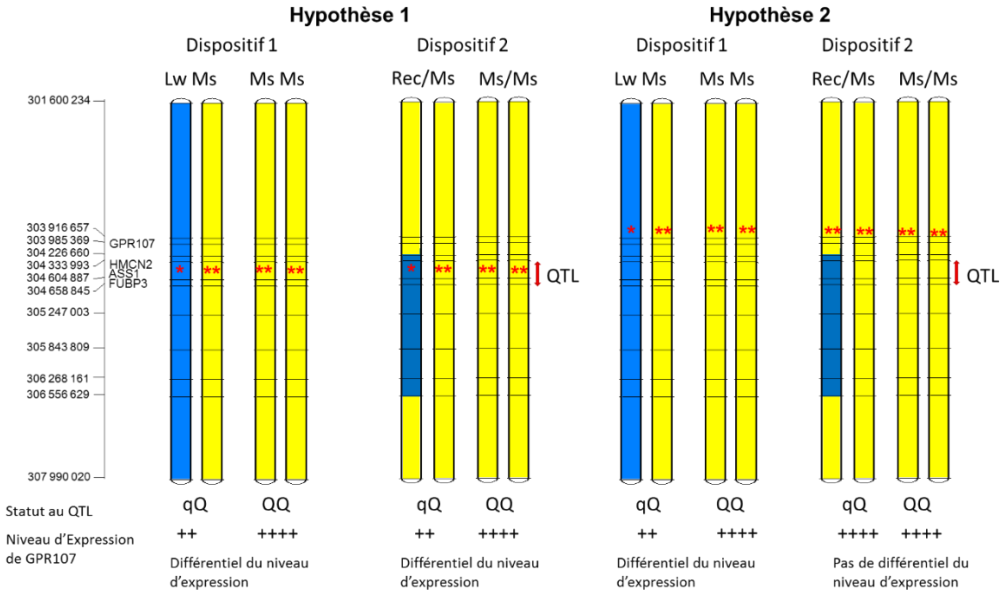


Figure 40 : Représentation des résultats attendus suivant les 2 hypothèses.

Pour les 2 dispositifs, les 2 classes de génotypes utilisés pour l'évaluation des deux hypothèses sont représentées selon la même nomenclature que la Figure 39 ; sous chaque classe le statut au QTL est rapporté. Les étoiles rouges indiquent la position théorique (selon l'hypothèse) du variant d'expression responsable de la différence de transcription (: allèle Lw; **: allèle Ms induisant une surexpression) ; les + schématisent le niveau d'expression théorique global du gène GPR107 (en fonction du génotype au variant d'expression).*

Compte tenu du coût de production des animaux, pour ce second dispositif des prélèvements des 3 tissus (foie, TASC, longe) ont été réalisés uniquement pour des animaux de 30 kg. Les prélèvements, préparations des échantillons d'ARN et cDNA ainsi que les analyses d'expression ont été réalisées dans les mêmes conditions que pour le premier dispositif. Néanmoins, les dispositifs étant totalement indépendants, les niveaux d'expression obtenus à partir des deux dispositifs ne peuvent être théoriquement comparés. Nous avons donc choisi de ré-analyser un sous-ensemble d'individus de chaque dispositif qui permettait une comparaison avec l'autre dispositif : les individus MsMs ont été comparés aux LwMs dans le dispositif 1, et les MsMS aux RecMs dans le dispositif 2.

Les nouveaux résultats obtenus sont présentés dans le Tableau 11. Dans le premier dispositif une différence de niveau d'expression significative pour GPR107 avait été obtenue dans la longe (p-value < 0,0007) et le foie (p-value < 0,004). Pour le second dispositif, un effet suggestif a été obtenu dans la longe (p-value < 0.06) et un différentiel significatif a été observé dans le TASC (p-value < 0,02) et le foie (p-value < 0.003). Cependant, les valeurs de significativité obtenues sont plus faibles que dans le premier dispositif et les valeurs de l'effet observé dans le TASC ne présentent pas le même signe. En effet, dans le dispositif 2 le niveau d'expression des individus MsMs est inférieur au niveau d'expression des individus RecMs (Effet négatif). En dehors du foie, les résultats obtenus dans les deux dispositifs ne sont pas comparables. Mais les valeurs de p-value obtenues pour les échantillons de TASC et de longe du second dispositif ne nous permettent pas de conclure avec assurance à l'existence ou non d'un différentiel d'expression entre les deux lots d'individus. Un nombre plus important d'individus serait nécessaire pour augmenter la puissance de cette seconde analyse.

Tableau 11 : Résultat des analyses de différentiel d'expression de GPR107 dans les 2 dispositifs expérimentaux mis en place, pour les 3 tissus et pour le stade à 30 kg.

	Dispositif 1				Dispositif 2			
	Effectif MsMs	**Effectif LwMs**	**Effet**	**p-value**	**Effectif MsMs**	**Effectif RecMs**	**Effet**	**p-value**
TASC	10	10	1.93	0.123	11	11	-3.64	0.026
Longe	10	10	1.37	0.00070	11	11	1.80	0.068
Foie	10	10	1.30	0.0042	11	11	3.82	0.0031

La même analyse de différentiel d'expression a été réalisée pour les gènes HMCN2, ASS1 et FUBP3 (Tableau 12). La significativité d'une différence d'expression entre les lots d'individus MsMs et RecMs pour ces différents gènes est comprise entre 2% et 6% et le signe des effets est négatif, indiquant que la présence d'un second chromosome Ms diminue le niveau d'expression de ces gènes.

Tableau 12 : Résultat des analyses de différentiel d'expression des 3 autres gènes présents dans l'intervalle dans les 2 dispositifs expérimentaux, pour le tissu adipeux et pour le stade à 30 kg.

	Dispositif 1				Dispositif 2			
	Effectif MsMs	**Effectif LwMs**	**Effet**	**p-value**	**Effectif MsMs**	**Effectif RecMs**	**Effet**	**p-value**
HMCN2	10	10	-1.80	0.431	10	10	-0.11	0.018
ASS1	10	10	0.03	0.753	10	10	-0.63	0.029
FUBP3	10	10	-1.40	0.104	10	10	-0.40	0.059

La mise en place de ce second dispositif ne nous a malheureusement pas permis de confirmer ou d'infirmer avec assurance que la mutation recherchée pouvait influencer l'expression des gènes localisés dans ou à proximité (GPR107) de l'intervalle de localisation du QTL.

6 ANALYSE IN-SILICO DES 3 GENES PRESENT DANS L'INTERVALLE

En parallèle des études d'expression une recherche bibliographique pour les 3 gènes restants dans l'intervalle a été réalisée afin d'évaluer si l'un d'entre eux pouvait présenter un lien fonctionnel avec le caractère étudié chez le porc. J'ai réalisé cette recherche principalement dans 2 bases d'annotation : AceView (Thierry-Mieg and Thierry-Mieg, 2006) et Genecards (http://www.genecards.org/) (Tableau 13).

Tableau 13 : Synthèse des données fonctionnelles issues des 2 bases d'annotation (AceView et Genecards).

Acronyme	NOM Gene	Taille	Nb transcrit	Nb Exons Humain	Process Biologique	Fonction	Domaine protéique
HMCN2	Hemicentin 2	168,2 kb	6 variants	98 exons	Bioluminescence	calcium ion binding	Immunoglobulin domain, Immunoglobulin I-set domain, Immunoglobulin V-set domain
ASS1	Argininosuccinate synthase 1	56.35 kb	10 Variants	16 exons	Arginine biosynthetic process	argininosuccinate synthase activity	Arginosuccinate synthase, A transmembrane domain
FUBP3	Far upstream element (FUSE) binding protein 3	58.79 kb	11 variants	22 exons	Regulation of transcription	protein binding	KH domain, A second peroximal domain

Le gène ***HMCN2*** (*Hemicentin 2*) est impliqué dans des voies du métabolisme du calcium et notamment dans la liaison avec les ions calcium. Parallèlement, il est décrit dans la littérature que les récepteurs couplés aux protéines-G peuvent être activés sous l'influence d'ions calcium, de ce fait, il est donc possible d'imaginer un lien entre ce gène et GPR107 (NCBI, Aceview).

Le gène **ASS1** (*Argininosuccinate synthase 1*) : La protéine codée par ce gène catalyse l'avant-dernière étape de la voie de biosynthèse de l'arginine. Dans le génome humain de très nombreuses copies de ce gène et de pseudo-gènes existent mais il semblerait que seule la copie localisée sur le chromosome 9 (orthologue de la copie présente dans la région du QTL) soit active pour la synthèse de l'argininosuccinate. Chez l'homme une mutation dans ce gène est à l'origine d'une maladie autosomique récessive, la citrullinémie. C'est un trouble rare du cycle de l'urée. Sur le plan clinique elle est caractérisée par une léthargie progressive, un refus alimentaire et des vomissements (http://www.orpha.net)

Enfin, **FUBP3** (*Far upstream element (FUSE) binding protein 3*) joue un rôle dans la régulation de la transcription, il pourrait donc avoir un lien avec l'ensemble des gènes de la région (NCBI, Aceview).

Finalement, au vu des données bibliographiques disponibles et des résultats d'expression différentielle aucun des trois derniers gènes présents dans l'intervalle ne semble avoir un lien direct avec le caractère étudié. Nous avons donc décidé de reprendre une étude exhaustive sur l'ensemble de l'intervalle et d'inventorier tous les SNP présents dans la région.

7 RESEQUENÇAGE COMPLET DES DEUX DERNIERS INDIVIDUS RECOMBINANTS

Pour réaliser la détection de tous les variants présents dans la région QTL du chromosome, nous avons procédé au séquençage de l'ensemble du génome des deux animaux recombinants à l'aide de la technologie HiSeq. Cette technologie permet d'obtenir 90 Gb de séquences de 2x 150 pb pour une ligne de séquençage sur un séquenceur HiSeq 3000. Afin d'ordonner ces séquences les unes par rapport aux autres, la séquence de référence du génome du porc est alors utilisée comme matrice pour procéder à l'alignement des

différentes lectures. Cette approche très puissante nécessite néanmoins de disposer d'une séquence de référence de bonne qualité. La région de l'intervalle de localisation du QTL n'étant pas couverte de manière continue par le draft Sscrofa10.2, nous avons cherché à améliorer dans un premier temps la séquence de référence porcine.

7.1 Reconstruction d'une nouvelle séquence de référence

Afin de reconstruire une séquence porcine continue sur l'ensemble de cette région, j'ai décidé d'utiliser les résultats obtenus précédemment lors du développement de marqueurs complémentaires (Partie 4 :_Recherche de l'intervalle minimum de localisation du QTL parmi les 6 clones BAC retenus à l'aide de l'outil CloneFinder, 2 clones suffisent pour maintenant couvrir l'intervalle de 380 kb : CH242-384I14 et CH242-487I19

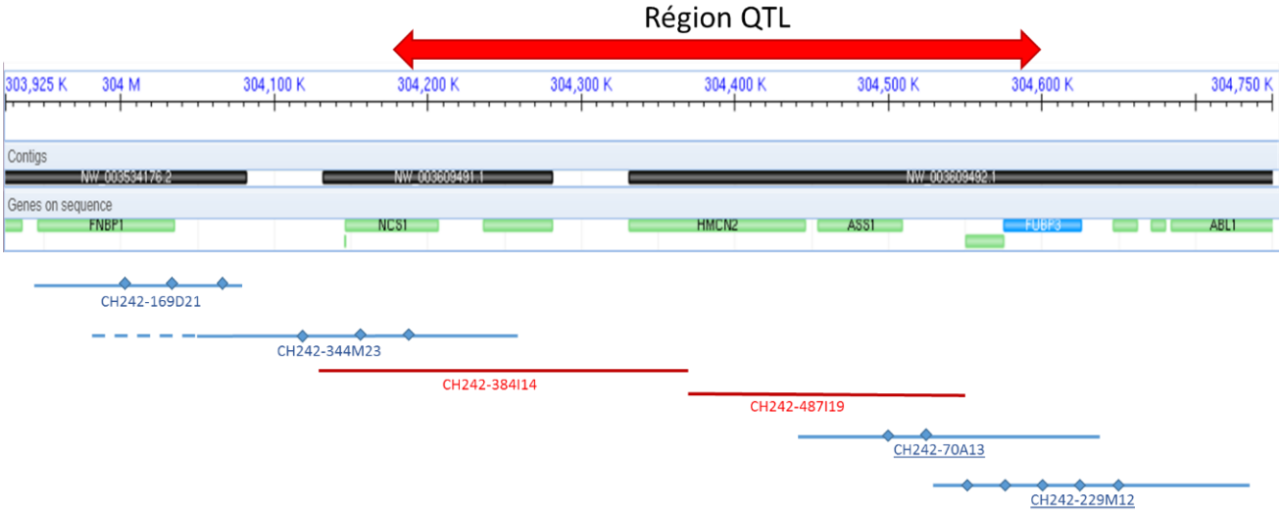


Figure 41 : contig de BAC couvrant l'intervalle de localisation du QTL de 673,4 kb localisé sur le chromosome 1 porcin.

Les BAC en bleu sont les clones qui ont été utilisés pour développer les marqueurs de type SNP, alors que les BAC en rouge ont été utilisés pour reconstruire la séquence de référence du nouvel intervalle QTL de 380 kb. Les contigs de la séquence de référence correspondent aux rectangle noirs, les zones blanches entre chaque contig représentent les régions sans séquences et représentées par une succession de 50 000 nucléotides inconnus (N).

J'ai donc comparé avec le programme Blastn (https://blast.ncbi.nlm.nih.gov/Blast.cgi) les séquences de chaque BAC avec la séquence humaine (version GRCh38.p10) homologue à la région QTL.

Le résultat de l'alignement du BAC CH242-384I14 sur la séquence humaine a permis d'identifier une homologie entre les positions 132 933 922 et 133 072 347 Mb du chromosome 9 humain et l'alignement du BAC CH242-487I19 entre les positions 133 069 117 et 133 420 017 Mb (Figure 42). Mais ces résultats ont également montré que la position de la fin du clone BAC 384I14 (133 072 347) se trouvait au-delà de la position du début du second BAC ce qui laissait supposer que ces 2 BAC étaient chevauchants.

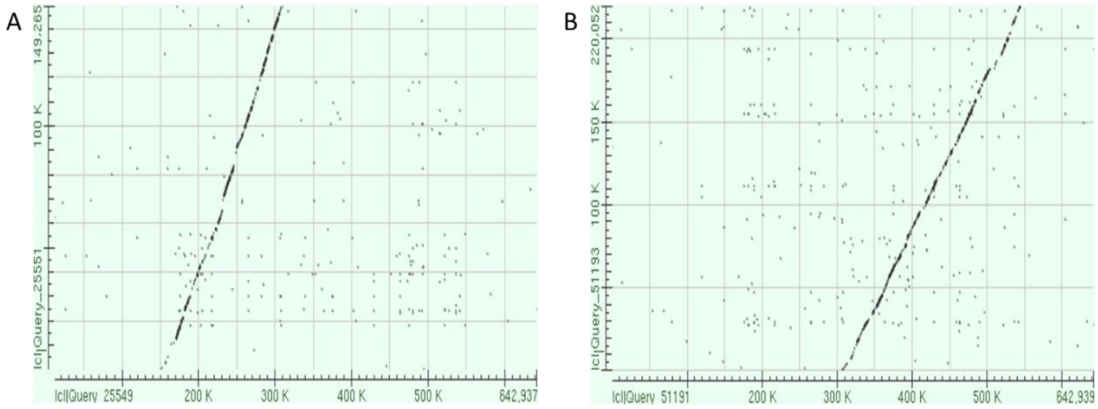


Figure 42 : Alignement des 2 clones BAC sur la séquence humaine homologue à la région QTL.
Représentation des points d'homologie entre le draft de référence de la séquence humaine (axe-X) et les séquences des BAC porcins (axe-Y).

Dans un second temps, j'ai donc réalisé la comparaison des extrémités des BAC porcins adjacents afin de rechercher une zone de recouvrement. En réalisant un Blast des 2 BAC l'un contre l'autre, nous avons mis en évidence une zone de chevauchement de 2 kb (entre les bases 147,266 et 149,265 du BAC CH242-384I14 et les bases 1 à 2000 du BAC CH242-487I19). Nous avons donc pu en conclure que malgré ce qui était décrit pour la séquence de référence, il n'y avait pas de séquence manquante pour cette région. Cette zone de chevauchement étant très petite, il est probable que les algorithmes automatiques d'assemblage n'aient pas retenu ce recouvrement.

Après avoir assemblé les séquences de ces deux BAC, nous avons remplacé la séquence du draft Sscrofa10.2 dans cette région par notre nouvelle référence afin de procéder à l'analyse des séquences obtenues pour les verrats recombinants.

7.2 Qualité des séquences obtenues

Chacun des deux verrats recombinants a été séquencé sur ¾ d'une ligne d'une flowcell d'un Run HiSeq3000 d'Illumina. Sur base de la capacité de ce séquenceur de 750Gb maximum/ flowcell (2 x 150 b en paired-ends), le nombre théorique de lectures que l'on devait obtenir était d'environ 234 375 000 par individu. A l'issue du séquençage le nombre total de séquences obtenues est respectivement de 421 681 592 (FR17MAG201003217) et 363 170 760 (FR18GAL200900504). Avant de procéder à l'alignement de ces séquences sur le génome de référence, un contrôle qualité des lectures a été réalisé. Ces contrôles sont réalisés par position au sein des lectures et par lecture. La Figure 43 résume les représentations graphiques des résultats de ces contrôles pour un des individus sur une des lignes.

- A l'échelle de la position, 3 critères sont estimés :
  - Un critère de qualité de la base (score calculé à l'aide l'algorithme de Phred (Ewing et al, 1998). Le score de qualité Q (Q = -10log10(e)) exprime le risque que le calling de la base soit faux. Des valeurs de Q de 10, 20 et 30 représentent respectivement un taux de précision de 90%, 99% et 99.9% (Figure 43A).
  - Le nombre de N (base non interprétée) (Figure 43B).
  - Le nombre de chacun des 4 nucléotides en chaque position : les pourcentages de nucléotides A et T doivent être équivalents de même que les pourcentages de C et G. L'écart entre les taux de AT et CG ne doit pas dépasser 20% (Figure 43C).

- A l'échelle de la séquence 6 autres critères sont estimés :
  - La qualité de la lecture estimée avec les mêmes valeurs de Q de l'algorithme Phred (Figure 43D).
  - La distribution de la taille des lectures, sachant que la longueur des lectures attendue est de 150 pb Figure 43E).
  - Le pourcentage moyen de GC au sein des lectures ne doit pas dévier de la distribution théorique pour plus de 30% des lectures (Figure 43F).
  - Le taux de duplication, permettant d'estimer le pourcentage de lectures non-uniques (Figure 43G).
  - L'identification de profil de K-mer, motifs courts sur-représentés (non représenté sur la figure)
  - La recherche de séquences sur-représentées (non représenté sur la figure).

Ces contrôles qualités sont réalisés systématiquement à la fin de chaque run et un rapport est fourni avec les données de séquençage sur la plateforme NG6.

De manière générale, la qualité des séquences obtenues pour nos 2 échantillons est très bonne. Pour les différents critères estimés, les « seuils d'alerte » n'ont pas été atteints.

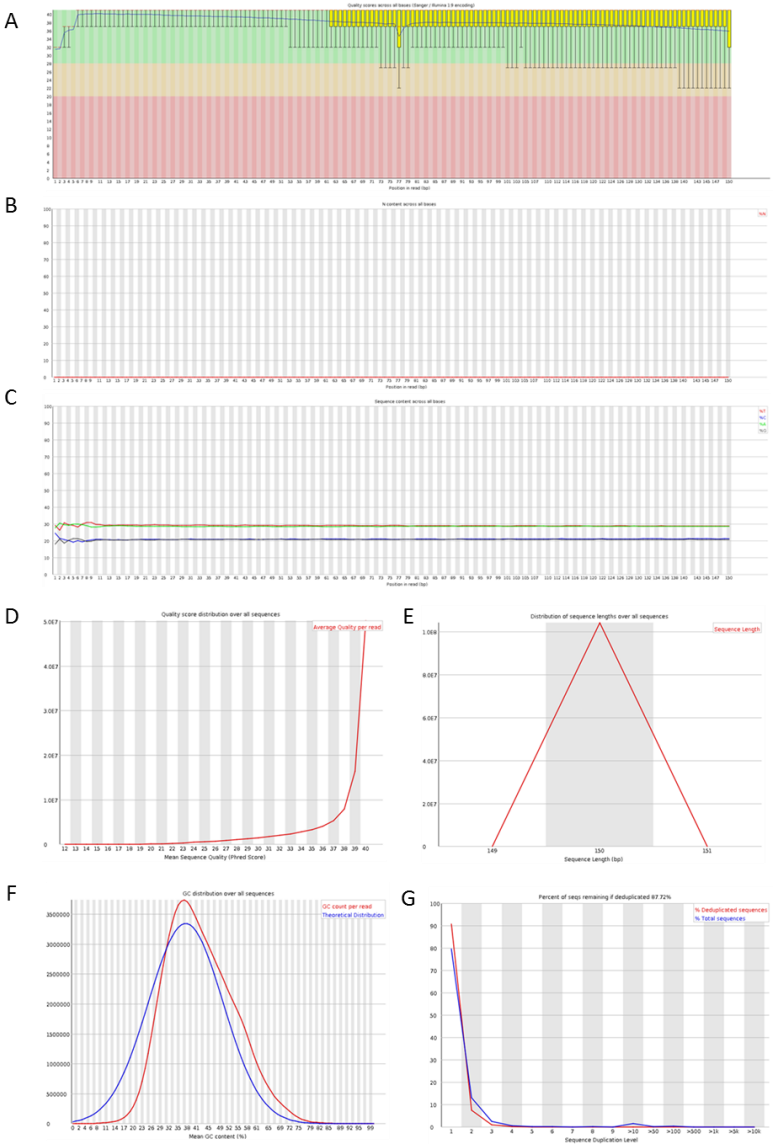


Figure 43 : Représentation du score de qualité le long de la séquence.

Les boîtes jaunes correspondent à la répartition des valeurs inter-quartiles (25-75%). Le haut et le bas des moustaches représentent les points à respectivement 90% et 10%. La ligne bleue représente la qualité moyenne par position. Les différentes parties du graphique (vert, orange, rouge) représentent respectivement des qualités très bonnes, raisonnablement bonnes et faibles.

Une fois la qualité des runs de séquençage validée, nous avons réalisé l'alignement des séquences pour chacun des 2 individus, à l'aide de l'outil BWA (Li *et al*., 2009), d'une part sur le génome de référence (Sus scrofa v10.2) et d'autre part sur le génome pour lequel nous avions corrigé la séquence dans l'intervalle du QTL. Les résultats de ces 2 alignements ont montré que l'alignement sur le génome corrigé a permis d'aligner environ 5000 lectures supplémentaires pour chaque individu (Tableau 14).

Tableau 14 : Comparaison du nombre de lectures du chromosome 1 qui s'alignent sur le génome de référence ou sur le génome de référence corrigé.

	Génome de référence v10.2	Génome de référence corrigé
FR17MAG2010003217	41 884 355 lectures	41 889 714 lectures
		+ 5187 lectures
FR18GAL20090504	35 989 514 lectures	35 995 320 lectures
		+ 5319 lectures

Un autre facteur qui permet de s'assurer que le variant détecté sera fiable est la profondeur de couverture pour chaque base, c'est-à-dire le nombre de fois que la base est lue. Harismendy *et al*., 2009 estiment que pour un fragment d'ADN séquencé 10 fois (couverture 10x) avec un score de qualité de séquençage supérieur à 20, la probabilité d'erreur est proche de zéro.

Compte tenu du nombre de lectures obtenues pour chacun des individus, le taux de couverture théorique sur l'ensemble du génome est de 21X (FR17MAG201003217) et 18X (FR18GAL200900504). Ce taux théorique a été estimé en prenant en compte le nombre de lectures obtenues multiplié par la taille de chaque lecture et divisé par la taille moyenne du génome.

Après alignement, la profondeur moyenne, pour notre région d'intérêt, est de 18X pour l'individu FR17MAG201003217 et de 15,5X pour l'individu FR18GAL200900504. Bien que ces valeurs soient un peu plus faibles qu'attendues, elles restent supérieures aux valeurs recommandées. Les variants qui seront alors détectés dans la suite des analyses ont donc une probabilité forte d'être de vrais polymorphismes.

7.3 Détection des variants

Après avoir validé la qualité des données, la recherche des variants a été faite grâce à l'outil GATK (McKenna *et al*., 2010) en observant les différences entre les lectures des 2 individus recombinants et la séquence de référence reconstruite.

A l'issue de cette analyse, le fichier de sortie au format VCF (Variant Call Format) présente l'ensemble des polymorphismes identifiés, selon leur position sur les chromosomes, l'allèle de référence et l'allèle alternatif, ainsi que des valeurs de qualité d'attribution du génotype. La valeur « QUAL » indique la probabilité que le polymorphisme existe vraiment. Cette valeur dépend de la qualité de la base et de la profondeur d'alignement. L'ensemble des résultats (lectures issues des données de séquençage HiSeq, séquence de référence, polymorphismes identifiés, gènes annotés) ont pu être ensuite visualisés grâce à l'outil IGV (Integrative Genomics Viewer) (Robinson *et al*., 2012; Thorvaldsdóttir *et al*., 2013). Dans un souci de simplification et de rapidité d'exécution du logiciel IGV, seul un intervalle de 1Mb comprenant l'intervalle du QTL a été extrait des fichiers de séquence.

7.3.1 Réduction de l'intervalle de localisation

L'interprétation de ce fichier et la visualisation des SNP dans IGV a permis tout d'abord de réduire l'intervalle de localisation.

En effet, le père FR18GAL20090504, verrat délimitant la borne haute de la région QTL, est homozygote pour plusieurs marqueurs consécutifs au-delà du marqueur MARC0031044 (position 304 276 686 pb), alors que ces mêmes marqueurs sont hétérozygotes pour le second père (Figure 44). Ce résultat a donc permis de réduire la zone du QTL de 27,989 kb au niveau de cette borne.

De la même façon, j'ai pu conclure que la zone d'homozygotie pour le père FR17MAG201003217, débutait à la position 304 597 687, alors que la position génétique du dernier marqueur de la borne basse était localisée à la position 304 605 149. Cela nous a donc permis également d'exclure 7, 462 kb de l'intervalle au niveau de la borne basse.

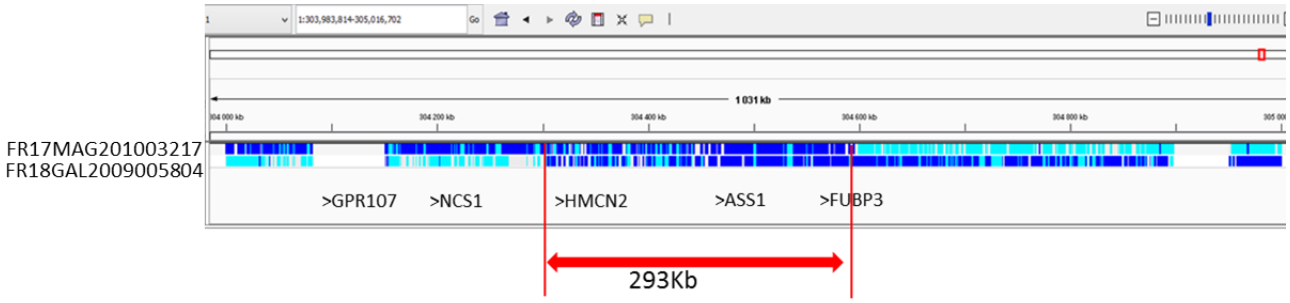


Figure 44 : Représentation de l'informativité des variants détectés pour les 2 pères recombinants.
Chaque ligne représente un individu et chaque barre un SNP, en bleu turquoise sont représentés les SNP homozygotes et en bleu marine les SNP hétérozygotes pour chacun des individus.

Au total, cela nous a permis d'exclure 35,4 kb de séquence et de réduire à nouveau l'intervalle du QTL à une région de 293 kb. Cette nouvelle zone comprend toujours les 3 gènes : HMCN2, ASS1, et FUBP3.

7.3.2 Bilan des variants détectés

Dans ce nouvel intervalle de 293 kb, la comparaison des deux individus recombinants par rapport au génome de l'animal Duroc utilisé comme référence a permis de détecter 4540 variants. Parmi ces polymorphismes certains sont des marqueurs dont un allèle est observé chez l'animal Duroc, l'autre allèle à l'état homozygote chez 1 ou les 2 verrats recombinants. Ces marqueurs ne sont pas de bons candidats de la mutation recherchée, puisque les deux verrats séquencés étant hétérozygotes au QTL, pour qu'un variant soit considéré comme mutation candidate, il est nécessaire qu'il soit hétérozygote pour les 2 pères ; par conséquent un premier filtre a été appliqué afin de ne conserver que les SNP hétérozygotes chez chacun des verrats. Parmi les 4540 marqueurs identifiés, 3052 variants étaient hétérozygotes pour le père FR17MAG201003217 et 3066 pour le second père. En appliquant cette condition nous avons pu réduire le nombre de SNP à 2395 variants soit 52,75 % de la liste initiale.

Cette liste encore trop importante ne permet pas d'envisager une analyse fonctionnelle de ces variants un à un, nous avons donc choisi de poursuivre cette étude par une analyse *in silico* de ces polymorphismes.

7.3.3 Annotation des variants détectés

L'étude de l'effet fonctionnel de ces 2395 variants a été réalisée avec le logiciel Variant Effect Predictor (VEP) sur le site Ensembl (http://www.ensembl.org/Tools/VEP).

VEP est un outil qui prend en charge tous les types de variants, les SNPs, les indels, les CNVs et les altérations de structure. Il combine les résultats de trois algorithmes publiés : PolyPhen (Adzhubei *et al*., 2010), SIFT (Kumar *et al*., 2009) et Condel (González-Pérez and López-Bigas, 2011). A partir des coordonnées génomiques des variations, VEP permet d'associer à chaque SNP les données d'annotation fonctionnelles connues à cette position (régions codantes, en amont d'un site de transcription, régions régulatrices, ARN non codant) et d'obtenir des informations sur les conséquences probables de ces variations sur les protéines.

Dans un premier temps, nous avons réalisé cette analyse avec les informations de structure et d'annotation de la version v10.2 du génome de référence. Parmi les SNP candidats, 925 variants (38,6%) étaient déjà référencés dans les bases de données et 1470 nouveaux variants ont été détectés (61.4%). Sur le plan fonctionnel, 90% sont situés dans les régions introniques et intergéniques (Figure 45A), les 10% restant sont localisés en **amont** (upstream_gene_variant : 3%), **dans** (3%) et en **aval** des gènes de l'intervalle (downstream_gene_variant : 4%). Comme le nombre de variations obtenues est important, nous avons tout d'abord fait le choix de regarder les 3 % des variants localisés dans la phase codante des gènes qui pouvaient présenter un impact fonctionnel important (Figure 45B). Parmi ces SNP, les variants synonymes ou neutre qui n'induisent aucun changement de l'acide aminé en raison de la dégénérescence du code génétique peuvent être écartés ; *a contrario* les mutations induisant un changement d'acide aminé (mutation faux-sens), restent des candidats potentiels. L'impact supposé de ces variants dépendra de différents facteurs, comme la position de l'acide aminé au sein de la structure de la protéine ou la nature physico-chimique de l'acide aminé (remplacement d'un acide aminé hydrophobe par un acide aminé hydrophile par exemple). Le dernier type de mutation qui induit généralement des effets délétères importants sont les mutations non-sens. Ces mutations génèrent l'apparition d'un codon stop prématuré, et la synthèse d'une protéine tronquée généralement dégradée ou l'élimination de l'ARNm tronqué.

(A) (B)

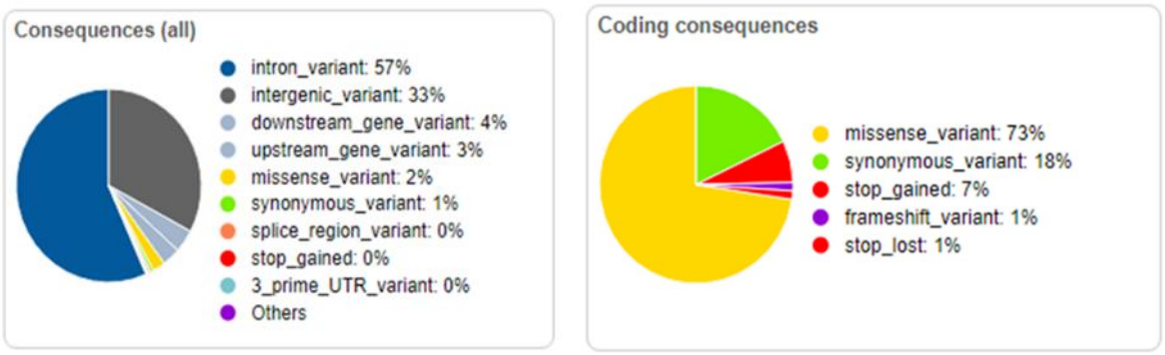


Figure 45 : Localisation (A) et impact fonctionnel (B) des 2395 variants suivant leurs positions sur la version 10.2 du génome de référence.

Les résultats obtenus montrent que 73% des mutations dans la phase codante des 3 gènes de l'intervalle sont identifiées comme des mutations faux sens (65 variant /89), 7 % comme des mutations non-sens (5/89) et seulement 18% (16/89) comme des mutations synonymes (Figure 45B). Ces résultats nous ont un peu surpris, le nombre de mutation faux-sens étant plus important que ce que l'on pouvait attendre. Or,

nous savions que la version 10.2 présentait des erreurs d'assemblage et nous pouvions donc également supposer que l'annotation de cette version n'était pas complète.

Au mois d'aout 2017 une nouvelle version du génome du porc (v11.1) a été publiée ainsi que les annotations fonctionnelles associées à ce draft. Nous avons réalisé un nouvel alignement sur cette version mise à jour, et une nouvelle analyse des variants identifiés.

Cette nouvelle version a confirmé que l'annotation de la version 10.2 était incomplète. Pour exemple, dans la version 10.2, le gène HMCN2 était décrit comme comportant 78 exons alors que chez l'Homme ce gène est composé de 98 exons. Dans la version v11.1 du génome, ce même gène comporte dorénavant 101 exons. A partir de cette nouvelle référence (v11.1), le nombre de variants hétérozygotes détectés pour les 2 pères dans l'intervalle du QTL de 293 kb est de 2429, donc très similaire au nombre détectés avec la version précédente (2395 variants). Parmi ces 2429 variants détectés, 1953 variants étaient déjà connus (80,4 %) et seuls 476 variants sont nouveaux (19,6%). Ces différences sont dues à l'augmentation du nombre de génomes du porc séquencés et publiés ces dernières années et utilisés pour produire le fichier de référence des variants connus chez le porc.

Les différences les plus notables portent sur l'impact fonctionnel des variants. Cette fois, 84 variants ont été détectés dans la phase codante des 3 gènes restants dans l'intervalle, 60 variants correspondent à des mutations synonymes, 22 variants sont des mutations faux-sens, 20 dans le gène HMCN2, 1 seul dans le gène ASS1 et 1 dans le gène FUBP3. Enfin, 2 variants ont été annotés comme pouvant avoir un effet délétère fort, en induisant l'apparition d'un codon Stop, 1 dans le gène HMCN2 et 1 dans le gène ASS1 (Figure 46B). Ces 2 variants nous ont donc paru très intéressants.

(A) (B)

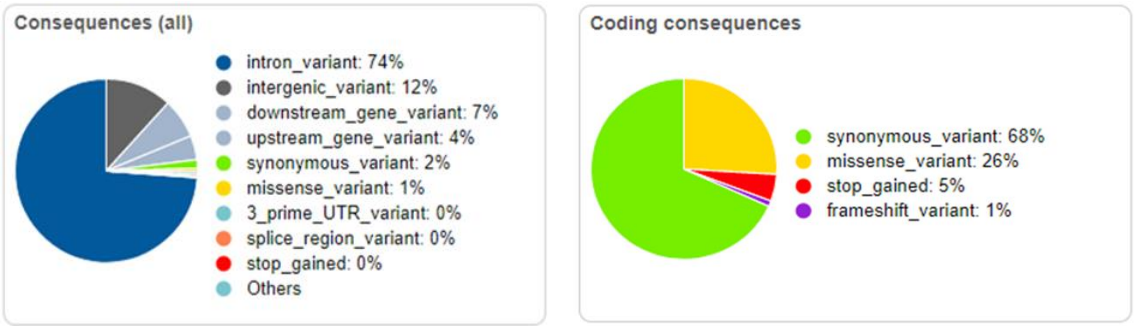


Figure 46 : Localisation (A) et impact fonctionnel (B) des 2429 variants suivant leurs positions sur la version 11 du génome de référence.

7.3.4 Analyse de 2 variants fortement délétères

7.3.4.1 Variant non-sens dans HMCN2

Nous nous sommes donc intéressés de plus près à ces 2 variants pouvant avoir un fort impact fonctionnel. Le premier variant est une insertion d'une séquence de 13 nucléotides à la fin de l'exon 91 du gène HMCN2. Généralement une insertion non multiple de trois bases entraîne au niveau des séquences codantes un décalage du cadre de lecture (*frame shift*) qui peut aboutir à l'apparition d'un codon stop ou modifier le site d'épissage. Tout d'abord, nous avons validé *in silico* la présence de ce codon stop, en comparant les 2 allèles présents chez les 2 pères recombinants.

- >All1_ref : CGTCATCCCTGAGAGCCTAGCGGACGCGGATCTGCAAGTGCAGgtggggtggacg
- >All 2_insertion +13pb : CGTCATCCCTGAGAGCCTAGCGGACGCGGATCTGCAAGTGCAGGTGGGGTGAGCAGgtggggtggacg

Figure 47 : Alignement et comparaison de la séquence de l'allèle 1 de référence avec l'allèle 2 porteur de l'insertion de +13pb dans l'exon 91 de HMCN2.

La séquence des exons est représentée en majuscule et en bleu alors que la séquence intronique est indiquée en minuscule et en noir. La séquence en rouge indique l'insertion de +13pb et en rouge et gras l'acide aminé TGA codant pour un codon stop. Les sites donneurs d'épissage sont indiqués en vert.

L'alignement des séquences a permis de mettre en évidence qu'une parfaite homologie entre les deux allèles est conservée en fin d'exon et qu'un site donneur d'épissage est re-créé à la position attendue (GT), en amont du codon stop. Cette structure laisse donc supposer que le codon stop pourrait être éliminé lors de l'épissage du transcrit porteur de l'insertion (Figure 47).

Afin de confirmer ce résultat, nous avons souhaité valider cette nouvelle hypothèse par PCR. Pour cela j'ai défini 2 couples d'amorces. Le premier a été choisi à partir de la séquence génomique, pour valider que le variant détecté n'était pas un artefact de séquençage ; le second couple a été quant à lui été défini pour amplifier, sur ADNc, un fragment encadrant l'insertion compris entre les exons 91 et 92.

La première PCR réalisé sur l'ADN génomique sur les individus du dispositif 2 a confirmé la présence de l'insertion. Les 2 allèles d'une différence de 13 pb étaient facilement identifiables sur gel d'agarose. Les génotypes obtenus avec ce marqueur coségrègent parfaitement avec l'allèle au QTL « d'épaisseur de lard dorsal ». La seconde amplification réalisée à partir de l'ADNc d'échantillons de tissu adipeux sous cutané de ces 41 mêmes animaux n'a pas permis de mettre en évidence une différence de taille entre les différents génotypes. Ce résultat corrobore l'hypothèse que le site d'épissage est maintenu malgré l'insertion et que les ARNm matures issus des deux allèles sont équivalents. Afin de valider totalement cette hypothèse, nous avons choisi de séquencer par la méthode de Sanger un individu homozygote pour l'allèle de référence et un individu homozygote porteur de l'insertion. Les séquences des 2 individus sont à 100 % identiques, cela confirme que l'épissage de l'intron 91 n'est pas modifié. La mutation n'est donc sans doute plus une mutation candidate pour le QTL impliqué dans l'épaisseur de lard dorsal, mais il n'y a pas eu pour le moment d'étude au niveau de la protéine.

7.3.4.2 Variant non-sens dans ASS1

Le second variant qui présentait un effet délétère fort est localisé dans l'exon 12 du gène ASS1. L'analyse des séquences des deux verrats a permis de mettre en évidence que deux bases adjacentes étaient variables. Si ces 2 SNP sont analysés indépendamment comme cela est réalisé par le logiciel d'annotation VEP, les résultats obtenus sont confirmés : la première mutation affecte la 1ème base du triplet d'un aa (TTG versus CTG), induit une mutation silencieuse et l'acide aminé n'est pas modifié (Figure 48B). La mutation en position 2 du triplet, induit un nouveau triplet (TTG vs TAG) et donc le codon stop mis en évidence par le logiciel VEP (Figure 48C). Lorsque les deux SNP sont pris en compte simultanément, le 1ier triplet TTG, correspond à l'allèle de référence, qui correspond à une Leucine et le second triplet CAG correspond à une Glutamine. Le codon stop précédemment détecté n'existe donc plus cependant cette mutation reste potentiellement une mutation candidate (Figure 48D), en effet sur le plan fonctionnel cela correspond à une mutation faux-sens.

(A)>ASS1_EX12_Allele_reference

ATC TAC GAG ACC CCA GCG GGG ACG ATT CTT TAC CAC GCT CAT TTA GAC ATC GAG GCC TTC
ACC ATG GAC CGG GAG GTG CGC AAA ATC AAA CAA GGC CTG GGC **TT**G AAA TTC GCC GAG CTG
GTG TAC ACG

IYETPAGTILYHAHLDIEAFTMDREVRKIKQGLG**L**KFAELVYT

(B)>ASS1_EX12_Allele_Alternatif_SNP1

ATC TAC GAG ACC CCA GCG GGG ACG ATT CTT TAC CAC GCT CAT TTA GAC ATC GAG GCC TTC
ACC ATG GAC CGG GAG GTG CGC AAA ATC AAA CAA GGC CTG GGC CTG AAA TTC GCC GAG CTG
GTG TAC ACG

IYETPAGTILYHAHLDIEAFTMDREVRKIKQGLGLKFAELVYT (mutation synonyme)

(C)>ASS1_EX12_Allele_Alternatif_SNP2

ATC TAC GAG ACC CCA GCG GGG ACG ATT CTT TAC CAC GCT CAT TTA GAC ATC GAG GCC TTC
ACC ATG GAC CGG GAG GTG CGC AAA ATC AAA CAA GGC CTG GGC TAG AAA TTC GCC GAG CTG
GTG TAC ACG

IYETPAGTILYHAHLDIEAFTMDREVRKIKQGLG*KFAELVYT (mutation non-sens)

(D)>ASS1_EX12_Allele_Alternatif_2SNP

ATC TAC GAG ACC CCA GCG GGG ACG ATT CTT TAC CAC GCT CAT TTA GAC ATC GAG GCC TTC
ACC ATG GAC CGG GAG GTG CGC AAA ATC AAA CAA GGC CTG GGC CAG AAA TTC GCC GAG CTG
GTG TAC ACG

IYETPAGTILYHAHLDIEAFTMDREVRKIKQGLGQKFAELVYT (mutation faux-sens)

Figure 48 : Séquences génomiques des 4 variants alléliques possibles dans l'exon12 de ASS1 et leur traduction en séquence protéique.

En dessous des 4 séquences nucléotidiques est indiquée la séquence protéique traduite. Les différentes analyses sont indiquées en couleurs : en bleu sont indiquées les analyses automatiques réalisées par VEP (outil de prédiction de l'effet des variants), les SNP sont analysés indépendamment en fonction de leurs positions. En vert, l'analyse réalisée manuellement en prenant en compte simultanément les 2 SNP.

Ces résultats mettent en évidence que même si ces outils de prédiction permettent d'orienter les choix des variants à étudier en priorité avec une bonne efficacité, il est indispensable de vérifier (*in silico*) et de valider en conditions expérimentales les variants détectés par ces analyses bio-informatiques avant de démarrer des expériences de validation fonctionnelle qui sont généralement difficiles à mettre en place et coûteuses.

Dans notre cas, le nombre de SNP potentiellement candidats trop important et l'absence de gènes candidat nous ont conduits à envisager une autre approche complémentaire de génétique afin d'en réduire le nombre.

Pour cela, nous avons décidé de réaliser le séquençage de la région d'intérêt sur un plus grand nombre d'individus pour lequel on avait connaissance de leur statut au QTL afin de réduire le nombre de variants partagés entre tous les individus.

8 SEQUENÇAGE D'UNE REGION CIBLEE DE 300 KB CHEZ PLUSIEURS INDIVIDUS HETEROZYGOTES AU QTL

8.1 Choix des animaux

Nous avons recherché l'ensemble des individus pour lesquels un testage sur descendance avait été réalisé dans la région du QTL du chromosome 1. L'objectif était de rechercher des verrats complémentaires, de génotype Q/q au QTL. Deux types de verrats ont été ainsi sélectionnés :

- Les individus BC du dispositif QTL : ces animaux sont tous des animaux composites Lw et Ms et nous faisons l'hypothèse que tous les verrats Q/q pour ce QTL seront porteurs de la même mutation et donc hétérozygotes à la mutation recherchée. Parmi l'ensemble des verrats testés, 11 sont hétérozygotes au QTL.

- Les individus issus du dispositif BIOMARK : le programme BIOMARK était un programme ANR destiné à évaluer si les QTL identifiés à partir du dispositif PorQTL étaient en ségrégation dans d'autres races et lignées françaises. Des familles composées d'un père et d'une centaine de descendants ont été constituées, phénotypées pour différents caractères de production (dont les épaisseurs de lard dorsal) et génotypées à l'aide de marqueurs choisis dans les régions des QTL. Dans la région du QTL du chromosome 1, le testage sur descendance de ces verrats nous a permis d'identifier 8 verrats hétérozygotes.

Le séquençage des deux verrats recombinants définissant l'intervalle minimum de localisation du QTL a permis de définir une première liste de polymorphismes candidats, correspondant à l'ensemble des SNP hétérozygotes chez ces deux individus (partie 7.3). Le séquençage d'autres verrats hétérozygotes (Q/q) devrait permettre de réduire la liste des variants hétérozygotes chez l'ensemble des individus. Au total 19 verrats complémentaires pourront être séquencés. Un schéma de cette stratégie est représenté Figure 49.

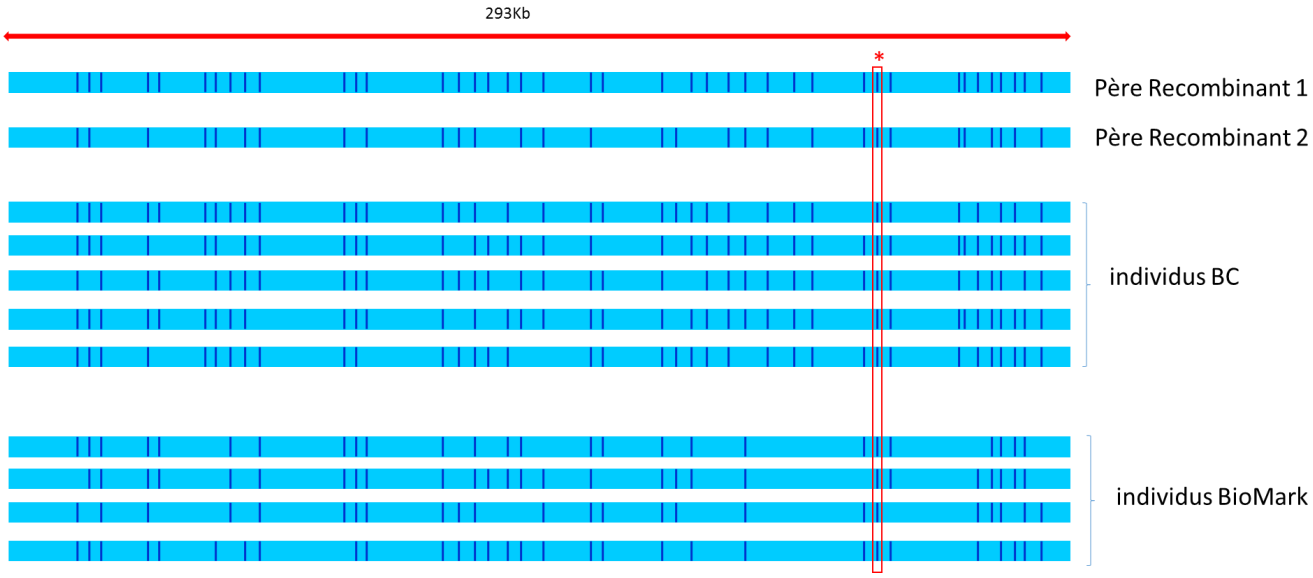


Figure 49 : Représentation de la stratégie pour la réduction du nombre de variants par séquençage d'amplicons de la région de 300 kb.

Les régions en bleu turquoise représentent les zones d'homozygoties et les variants sont représentés par des barres verticales en bleu marine. La mutation candidate présente chez tous les individus est symbolisée par une petite étoile rouge et un rectangle rouge.

Nous avons dans un premier temps cherché à reconstruire les haplotypes pour les 19 animaux sélectionnés. Dans la région du QTL, les individus BC (LwxMs) portent tous un haplotype Ms, identique à celui porté par les deux verrats recombinants, et un haplotype Lw. Contrairement aux chromosomes Ms, les 11 chromosomes Lw présentent une forte variabilité et devraient permettre de réduire le nombre de variants commun entre tous les individus.

Les 8 verrats du programme BIOMARK présentent une plus forte variabilité haplotypique en raison de la diversité des populations dont ils sont issus. Parmi eux, un individu semble particulièrement intéressant, l'animal 18GAL030759. En effet, cet individu présente deux haplotypes d'origine Lw très comparables dans la région du QTL. Parmi les 15 SNP (présents sur la puce 60K) de la région, cet individu présente un génotype homozygote à l'exception du marqueur SNP ALGA0010839.

Afin de confirmer que l'individu 18GAL030759 est peu variable dans la région du QTL, nous avons développé 7 couples PCR de 1000 pb en moyenne. Six couples ont été choisis tous les 50 kb dans l'intervalle du QTL de 300 kb afin de couvrir l'ensemble de la région et 1 couple a été choisi autour du SNP ALGA0010839. Nous avons fait le choix de séquencer des fragments d'assez grande taille pour être avoir plus de chances de détecter au moins 1 SNP dans chaque fragment. En effet la fréquence moyenne d'un SNP dans l'espèce

porcine est de 1 tous les 300 pb. Nous avons réalisé le génotypage par séquençage de cet individu, ainsi que le père recombinant FR18GAL200900504, comme individu témoin.

L'individu 18GAL030759 n'est variable que dans 2 des 7 amplicons testés, la zone en amont de HMCN2 et la zone qui comprend le marqueur SNP ALGA0010839 (Figure 50). Enfin, parmi les 86 SNP détectés seulement 7 sont partagés avec l'individu recombinant FR18GAL200900504, soit moins de 10%.

Figure 50 : Résultats du séquençage des 7 régions de 1 kb de l'individu 18GAL030759.

Par conséquent ce résultat très encourageant laisse supposer que cet individu ne présente que de très petites régions d'hétérozygotie, au moins 2, et qu'il devrait donc nous permettre d'éliminer un grand nombre de variants détectés lors de l'approche précédente.

8.2 Choix de la stratégie mise en place

Même si l'approche par séquençage du génome entier reste la stratégie la plus exhaustive et la plus fiable car elle n'introduit pas des biais d'amplifications, elle reste cependant une méthode coûteuse et ne peut pas être envisagée lorsqu'on souhaite séquencer un grand nombre d'individus.

Pour séquencer l'intervalle de 293 kb de la région QTL chez les 19 verrats hétérozygotes au QTL du chromosome 1, nous avons choisi une méthode de séquençage d'amplicons. Pour cela nous avons défini 30 couples de 10 kb pour couvrir la zone. L'amplification de ces 30 couples de 10 kb a été réalisée avec la PrimeSTAR® GXL de chez Takara selon les conditions décrites dans la partie Matériels et Méthodes en s'efforçant de n'avoir qu'une seule condition PCR pour les 30 couples.

Les 30 amplicons de chaque individu ont été ensuite poolés et une librairie de séquençage a été réalisée pour chaque individu. Chaque librairie étant marquée par un adaptateur différent, plusieurs individus ont été mélangés pour réaliser le séquençage. A ce jour, la totalité des amplicons ont été obtenus et dosés pour l'ensemble des animaux et 8 premiers verrats, dont l'individu 18GAL03759, sont en cours de séquençage.

Si les premiers résultats sur les 8 individus confirment que la stratégie développée et les différentes étapes du protocole sont bien maitrisées, cette même approche sera alors réalisée sur les 11 autres individus hétérozygotes au QTL. Enfin, il est fort probable que l'analyse de tous ces individus et notamment de l'individu 18GAL030759, qui présente plusieurs régions d'homozygotie, nous permette de réduire le nombre de variants à une centaine SNP candidats.

DISCUSSION- PERSPECTIVES

CHAPITRE IV : PERSPECTIVES – DISCUSSION

Le principal objectif de ce projet était d'identifier la mutation causale d'un QTL localisé sur le chromosome 1, responsable d'une part de la variabilité de l'épaisseur de lard dorsal chez le porc. Pour cela, j'ai réalisé plusieurs approches indépendantes et complémentaires (cartographie fine, transcriptomique et séquençage haut-débit). Même si aujourd'hui, ces différentes approches ne nous ont pas encore permis d'identifier la mutation causale, j'ai pu d'une part réduire de façon significative la taille de l'intervalle de localisation du QTL, de 1,6Mb à 280 kb et d'autre part, grâce au séquençage des 2 individus recombinants qui définissaient les bornes haute et basse de cet intervalle, j'ai pu référencer l'ensemble des polymorphismes candidats. A l'issue de ces résultats, aucun gène de l'intervalle, sur la base des connaissances disponibles, ne semble être un bon candidat fonctionnel et le nombre de polymorphismes est encore trop important pour pouvoir envisager directement des approches de validation fonctionnelle. D'autres analyses sont donc en cours pour continuer à diminuer ce nombre de variants candidats.

1 PERSPECTIVES DES RESULTATS A OBTENIR A COURT TERME

1.1 Réduction du nombre de variants et identification de la mutation causale

Au moment de la rédaction de ce manuscrit, nous avons entrepris de séquencer la région de 280 kb des 19 individus hétérozygotes au QTL. Grace à ces données de séquençage, nous pensons pouvoir fortement diminuer le nombre de polymorphismes qui ségrégent parfaitement avec les allèles au QTL, notamment à l'aide d'un verrat de race sino-européenne (LwxMs) qui présente plusieurs régions homozygotes au sein de ces 280 kb. Cet individu devrait nous permettre de réduire de façon très significative le nombre de variants pour ne disposer plus que d'un petit nombre de candidats (une centaine).

Bien que réduire la liste des mutations candidates de 2395 variants à une centaine est une avancée significative, la mise en œuvre de tests fonctionnels de 100 mutations candidates reste inenvisageable. A court terme, afin de poursuivre l'étude génétique des marqueurs qui resteraient encore candidats, il est envisagé de tester l'effet de chacun des polymorphismes restants chez les familles du dispositif F2 PORQTL dans lequel ce QTL avait été initialement détecté. Après avoir génotypé l'ensemble du dispositif pour les différents polymorphismes candidats, le génotype des individus à chaque SNP sera successivement pris en compte comme effet fixe dans le modèle d'analyse destiné à estimer l'effet du QTL. Cette stratégie permettra d'identifier le marqueur ou les marqueurs qui expliqueront la plus grande part de la variabilité génétique du QTL. En effet, si en intégrant les génotypes d'un des marqueurs comme effet fixe dans le modèle, l'effet de la région sur le caractère épaisseur de lard dorsal n'est plus significatif, c'est que le SNP pris en compte est causale ou à minima en total déséquilibre de liaison avec la mutation.

En complément du dispositif PORQTL, cette approche sera également menée dans un second dispositif constitué d'un échantillonnage d'une population sino-européenne, la Taizumu. La lignée Taizumu, est une lignée composite qui a été créée en 1994, par l'organisme de sélection Axiom, en inséminant des truies chinoises Meishan avec de la semence de verrats hyper-prolifiques de race Large White. Après 3 générations sans sélection, les animaux ont été principalement sélectionnés sur des caractères de qualités maternelles mais aussi pour des caractères de croissance et de composition de carcasse. Dans le cadre de sa thèse, Maxime Banville avait démontré, par une étude d'association réalisée sur un millier d'individus, que cette région chromosomique du chromosome 1 influençait l'adiposité et le nombre de tétines des animaux

(Banville et al., 2015). Nous avions conclu de cette étude que la mutation que nous recherchions était également en ségrégation dans la population Taizumu.

A l'issue de l'analyse des 19 séquences complémentaires, les polymorphismes candidats seront donc génotypés sur ces deux dispositifs via la technologie Trueseq Genotype NE d'Illumina qui permet d'obtenir simultanément le génotype de quelques centaines d'animaux pour quelques centaines de SNP (Illumina, 2017). L'effet de chaque marqueur sur la variabilité du caractère pourra alors être testé afin de ne conserver que les polymorphismes candidats permettant d'expliquer la variabilité phénotypique déterminée par cette région chromosomique. Nous espérons ainsi ne conserver qu'un petit nombre de candidats que nous pourrons alors valider ou invalider via des approches fonctionnelles.

1.2 Validations fonctionnelles de la mutation causale

1.2.1 Analyse *in silico*

Si comme attendu les analyses de séquençage complémentaires nous permettent de réduire fortement le nombre de mutations, les premières analyses mises en œuvre seront des analyses *in silico*, comme cela a été réalisé précédemment à l'issue du séquençage des 2 individus recombinants.

Ces analyses *in silico* permettent de hiérarchiser la liste des polymorphismes candidats en identifiant leurs positions relativement aux éléments fonctionnels connus et annotés. Nous porterons une attention particulière aux polymorphismes délétères et aux polymorphismes localisés au niveau de sites d'accrochage de facteurs de transcription. Ces courtes séquences d'ADN de 6 à 30 nucléotides de long peuvent être retrouvées en amont ou en aval des promoteurs voire dans les régions introniques des gènes et ont un effet important sur le niveau de régulation des gènes.

Les mutations candidates seront alors considérées une à une et quelques analyses simples de bioinformatique seront faites afin de valider ou non leurs effets supposés. Pour rappel, l'analyse faite à l'issue du séquençage des 2 individus recombinants nous avait permis d'exclure 2 mutations qui avaient été prédites comme fortement délétères (mutation stop dans le gène HMCN2 et ASS1).

1.2.2 Validation des mutations candidates

Après prédiction *in silico*, si le nombre de mutations candidates est suffisamment réduit (à moins d'une dizaine), alors des approches de validation fonctionnelle pourront être envisagées. Ces approches seront destinées à prouver l'impact de la mutation, en recréant le phénotype muté dans des cultures cellulaires ou chez une espèce modèle. Le choix de la méthode de validation dépendra donc essentiellement de la position ou de la nature de la mutation causale identifiée. En effet si la mutation est localisée dans la région promotrice, son effet aura plus de chances d'impacter le niveau d'expression du transcrit du gène ; il convient donc en général de quantifier ce niveau d'expression par PCR-quantitative. Or dans notre étude cette approche a déjà été menée et ne nous a pas permis de mettre en évidence une différence d'expression en fonction du génotype. Cependant pour ces approches, le choix du ou des tissus à étudier est crucial, ainsi que l'âge de l'animal au moment du prélèvement. Certains gènes, par exemple, sont exprimés majoritairement pendant le développement embryonnaire, et l'étude de leur niveau d'expression après la naissance de l'animal ne permet pas de conclure.

Les autres méthodes qui peuvent être également mises en œuvre pour tester une variation de l'expression des gènes reposent sur des expériences d'expression *in vitro* de gènes rapporteurs, par exemple avec des constructions cellulaires qui portent les 2 formes du promoteur en amont du gène de la luciférase.

Enfin si la mutation implique une modification d'un site de fixation de facteur de transcription, nous pourrons également envisager des expériences de gel retard. Il est possible de déterminer la nature du facteur de transcription en utilisant des anticorps spécifiques. Si les profils de migration obtenus sont différents en fonction des génotypes, il sera possible de conclure à l'influence de la mutation candidate sur la capacité de fixation du facteur de transcription.

Dans le cas où la mutation aurait un effet supposé sur le niveau d'expression de la protéine il sera alors nécessaire, dans un premier temps, d'acquérir des connaissances sur le fonctionnement normal de la protéine non mutée (niveau d'expression et/ ou localisation) puis dans un second temps de comparer les 2 génotypes. L'estimation de la quantité de la protéine pourra se faire classiquement par Western-blot. Pour ce type d'expérience une difficulté en fonction du gène analysé restera sans doute l'accès à des anticorps qui puissent fonctionner chez l'espèce porcine.

Des informations de localisation de la protéine pourront également nous aiguiller sur une fonction probable de cette protéine : une mutation peut affecter la présence ou les conditions d'adressage, de transport, d'ancrage de la protéine à d'autres molécules. La localisation de la protéine pourra être étudiée par immunohistochimie ou immunofluorescence sur des coupes de tissus. Par exemple, l'hémicentine codé par le gène HMCN2, un des trois gènes encore candidat est décrite comme une protéine de la matrice extracellulaire impliquée dans le contact, l'adhésion et la migration cellulaire, on pourrait rechercher des différences de localisation de la protéine en fonction du génotype étudié. Toutefois, aucun lien direct de HMNCN2 n'a été mis en évidence pour l'heure avec le tissu adipeux.

1.2.3 Validation fonctionnelle au niveau cellulaire

L'utilisation de cultures de cellules permettrait de tester des mécanismes moléculaires comme une modification de la morphologie cellulaire ou de l'activité enzymatique. Les cellules seront préalablement transformées par mutagénèse dirigée, de sorte que les deux génotypes puissent une nouvelle fois être comparés.

La dernière option, la plus lourde à mettre en œuvre pour valider la mutation, consisterait à vouloir reproduire le phénotype dans une espèce modèle telle que la souris ou le poisson zèbre, le choix de l'espèce modèle dépendant du gène à étudier et donc de sa fonction. S'il y a quelques années cette stratégie était délicate à mettre en œuvre et souvent infructueuse en raison de difficultés techniques, cette option peut être aujourd'hui plus largement envisagée grâce à l'évolution des outils de modification du génome (CRISPR/Cas9).

2 DISCUSSION DES STRATEGIES UTILISEES

2.1 Stratégie de cartographie génétique

Comme évoqué dans la partie bibliographique, depuis la mise en place des premiers programmes de cartographie de QTL les approches génétiques utilisées ont grandement évolué. Initialement la stratégie mise en place pour identifier une mutation causale pouvait se résumer en 3 grandes étapes : la primo-localisation, la cartographie fine (approche de croisement en retour via des dispositifs familiaux ou par l'exploitation du déséquilibre de liaison dans la région candidate à partir des populations commerciales) et enfin l'identification d'un gène candidat et de la mutation causale. Les deux premières étapes se révèlent extrêmement couteuses

et longues pour certaines espèces et notamment chez les bovins et les porcins. Chez le porc, malgré de grandes tailles de portées (douze porcelets en moyenne), une durée de gestation de 114 jours et un intervalle de génération d'un an, la production d'animaux est relativement lente. A titre d'exemple, la production de 7 générations de BC pour la réalisation de la cartographie fine du QTL localisée sur le chromosome 1 a pris une douzaine d'années.

Au cours de ces dernières années, l'évolution des méthodes et des outils en génétique animale a révolutionné les projets de recherches, notamment avec le développement des puces de génotypages Haute Densité (commercialisation de la puce porcine en 2009 composée de 60 000 SNP). La densité en marqueurs est telle qu'il est maintenant possible de réaliser des études d'association tout génome, en exploitant le déséquilibre de liaison, dans l'ensemble des populations commerciales. Ces études, appelées GWAS, permettent donc de combiner les 2 premières étapes en une seule, et d'obtenir des intervalles de localisation tout aussi résolutifs que par les approches de cartographie fine plus classique avec un gain de temps considérable : une telle analyse peut être effectuée en moins d'un an.

Au vu de ces différents succès, actuellement, la cartographie fine par production de Backcross est abandonnée au profit des analyses d'association.

2.2 Stratégie de séquençage

Une fois l'intervalle de localisation identifié et limité à quelques centaines de kilobases, l'identification de la mutation causale passe par une étape de séquençage de la région. Aujourd'hui bien que séquencer un génome complet s'avère plus simple et plus rapide que de séquencer une région ciblée du génome, cette étape reste cependant encore très couteuse et se limite donc au séquençage de deux ou trois individus. Les individus choisis sont donc généralement un individu sain et un individu malade lorsque la mutation recherchée détermine une maladie, ou comme dans notre étude les 2 individus qui délimitent les bornes haute et basse de l'intervalle de localisation du QTL. Cependant, à l'issue de ce séquençage le nombre de variants candidats reste encore trop important. Pour réduire ce nombre, il est alors nécessaire d'augmenter le nombre d'individus à analyser.

Pour pouvoir réaliser le séquençage sur un plus grand nombre d'échantillons pour un coût raisonnable, il est donc indispensable de travailler sur une plus petite fraction du génome à séquencer. Pour cela et jusqu'à présent 2 stratégies de séquençage d'une région ciblée étaient principalement envisagées (Mertes et al., 2011) : (1) Une approche par capture de la région d'intérêt par hybridation sur support solide (puces à ADN) ou (2) par amplification individuelle de grands fragments (10 à 30 kb).

Le principe de la première méthodologie repose sur la capture par hybridation de l'ADN des régions cibles avant séquençage. Pour que cette technologie soit efficace il est indispensable de disposer d'une parfaite connaissance de la séquence de référence, afin que les sondes utilisées pour la capture représentent de façon exhaustive la région d'intérêt. Cette approche est très souvent utilisée en génétique humaine pour rechercher l'ensemble des polymorphismes présents dans l'ADN codant ; des kits de capture universelle des exons des gènes humains sont commercialisés (capture et séquençage d'exome) et permettent la recherche pan-génomique de mutations candidate exonique. Lorsque l'objectif est de cibler une région particulière du génome cette technologie nécessite la réalisation d'un support de capture spécifique ; cette approche n'est donc surtout financièrement avantageuse que lorsqu'un très grand nombre d'individus doit être séquencé.

La seconde technologie est la plus rapide à mettre en place et ne nécessite pas d'équipements particuliers. Elle consiste à obtenir la région d'intérêt à l'aide d'amplicons de grande taille. Ces amplicons peuvent aller jusqu'à 30 kb en fonction des enzymes utilisées et des régions à amplifier. L'ADN utilisé en

séquençage correspond alors au mélange des différents amplicons ciblant la région d'intérêt. Cependant, il est important d'avoir conscience que cette approche peut générer des problèmes d'homogénéité de la profondeur, qui dépend de l'efficacité de PCR de chacun des fragments. Un biais technique peut également résulter de la présence d'allèle nul (polymorphisme en 3' des amorces) induisant l'amplification d'un seul des deux allèles.

Récemment, une nouvelle méthode basée sur la technologie CRISPR-CAS9 (Jiang et al., 2015) a été développée. Elle permet de capturer la séquence d'intérêt sans passer par ces étapes d'amplification ou d'hybridation. Ceci permet contrairement aux 2 techniques précédentes de ne pas se limiter à des zones bien caractérisées du génome et de limiter les erreurs dues aux biais d'amplifications. En effet, il suffit de définir 2 courtes séquences d'ARN d'une vingtaine de nucléotide qui serviront de guide à la protéine Cas9. Cette protéine, qui appartient à la famille des nucléases agira comme un véritable ciseau moléculaire au niveau des séquences guides. Comme dans le cas de la capture de la région d'intérêt par amplification de grands fragments, cette méthodologie est relativement peu couteuse et simple à mettre en place. Cependant, une des difficultés majeures de cette technologie réside dans la préparation de l'ADN génomique. En effet, les molécules d'ADNg obtenues après extraction doivent être de très haut poids moléculaire. C'est pourquoi il est préconisé de réaliser l'extraction d'ADNg après l'avoir emprisonné dans des blocs d'agarose (plugs), pour conserver l'intégrité de la molécule d'ADN. Cette difficulté peut cependant se révéler un avantage, lorsque cette approche est couplée aux 3ieme générations de séquençage (séquencage sur molécule unique). En effet, lorsque le séquençage est directement effectué sur l'ADN génomique natif, il est alors possible de préserver la composition génétique et épigénétique de la molécule et donc de posséder des informations supplémentaires et complémentaires.

Dans notre cas, c'est la seconde méthode par amplification de grands fragments que nous avons retenue, car nous avions seulement une vingtaine d'échantillons, elle était donc financièrement la plus avantageuse et la plus simple à mettre en place. Cependant les couts de séquençage continuent de diminuer, et les compagnies engagées dans le développement de nouveaux séquenceurs promettent qu'à terme, le coût de séquençage d'un génome pourrait être de 100 $; il est donc légitime de se demander si ces stratégies de re-séquençage régional seront encore utilisées, car il sera alors plus facile, plus rapide et moins couteux de re-séquencer directement le génome de chaque individu en totalité.

3 INTERET DE TROUVER LA MUTATION CAUSALE

La recherche des régions chromosomiques, des gènes et des mutations influençant des caractères d'intérêt agronomique est un enjeu majeur pour les filières animales. En effet, le développement de la génétique moléculaire et de la génomique chez les espèces agronomiques, à partir des années 90, avait pour objectif de mettre en évidence les polymorphismes affectant des phénotypes d'intérêt afin de les prendre en compte dans les schémas de sélection. La situation idéale serait de pouvoir sélectionner les animaux directement sur les mutations causales, et de réaliser une sélection assistée par gènes (GAS : Gene Assisted Selection). Or ce type de sélection ne concerne que quelques gènes et principalement des gènes majeurs. En effet, au vu de la difficulté et du temps nécessaire à aboutir à cette identification, notamment pour des caractères quantitatifs, elle ne peut donc pas être appliquée à l'échelle du génome, toutes les mutations causales des QTL n'étant pas connues.

Depuis quelques années, suite à la commercialisation des puces haute-densité, une nouvelle approche de sélection sur marqueurs a pu être envisagée ; cette sélection est basée sur la prédiction d'un « score moléculaire ». Le principe de cette « sélection génomique » consiste à prédire la valeur génétique d'un individu à partir d'un réseau de marqueurs denses (SNP) couvrant l'ensemble du génome. L'hypothèse sous-

jacente de cette approche est que la majeure partie de la variance génétique est expliquée par de nombreux QTL (Quantitative Trait Loci), la plupart à petits effets ; si la densité de marqueurs est suffisamment élevée alors certains vont se retrouver en déséquilibre de liaison avec ces QTL, au moins intra-race. Au préalable les relations entre phénotypes et marqueurs doivent être définies à partir d'une population de référence suffisamment grande, constituée d'individus disposant à la fois des phénotypes d'intérêt et des génotypes aux marqueurs génétiques. Ces relations sont ensuite extrapolées pour prédire la valeur génétique des candidats à la sélection, sur la base de leur génotype aux marqueurs, bien avant qu'ils ne disposent de performance. La force de cette stratégie est de pouvoir intégrer les marqueurs dans une approche de sélection génomique, sans qu'il soit nécessaire d'avoir identifié et validé la mutation causale par des approches fonctionnelles. Le gain permis par la sélection génomique par rapport à la sélection pratiquée jusqu'à maintenant est d'autant plus important que les caractères sont peu héritables, difficiles à phénotyper, d'expression tardive, ou coûteux à mesurer (Tribout, 2013).

Même si cette sélection a d'ores et déjà montré son efficacité, notamment chez les bovins laitiers, les marqueurs ne permettent généralement pas de mimer totalement les mutations des QTL associés (Fritz et al., 2013). La raison principale est qu'en général le déséquilibre de liaison (DL) entre les marqueurs analysés et les QTL est rarement total, et que ce DL n'est souvent pas maintenu dans les différentes races. Chercher à identifier la mutation causale reste donc encore pertinent dans certains cas, comme par exemple pour contre-sélectionner des allèles défavorables de gènes à effet fort sur le caractère comme ce fut le cas pour les loci « RN » ou « HAL » associés à des défauts de qualité de viande ou dans le cas de gènes impliqués dans le déterminisme d'anomalies congénitales (Riquet et al, 2011.).

La sélection sera d'autant plus précise que les mutations causales des loci dont les effets sont les plus forts pourra être intégrée à la sélection génomique.

Dans le cadre de la cartographie du QTL d'engraissement du chromosome 1, nous avons fait le choix de mener des travaux destinés à identifier la mutation causale. L'effet de ce locus est important et présente un mode de ségrégation intéressant à considérer dans un schéma de sélection. Comme présenté précédemment l'allèle d'origine MS « de fort engraissement » est récessif face à l'allèle « maigre » d'origine LW. Alors que pour les porcs charcutiers, des animaux avec une forte teneur en viande maigre et une faible épaisseur de lard dorsal sont souhaités, ce caractère « d'engraissement » reste important pour la carrière des truies. En effet, une sélection forte pour la croissance maigre et une augmentation de la prolificité des truies a affecté la capacité de ces dernières à allouer leurs ressources dans les différentes fonctions biologiques. Les truies allaitantes disposent de réserves adipeuses plus faibles que par le passé et ont un appétit réduit avec comme conséquence directe une diminution de leur capacité à reconstituer des réserves entre deux mises bas. Il serait donc intéressant de pouvoir exploiter la récessivité de l'allèle Meishan induisant une épaisseur de lard dorsal plus importante, en produisant des truies homozygotes pour cet allèle. Utilisées en croisement avec des verrats homozygotes pour l'allèle « maigre », les produits (porcs charcutiers) seront alors maigres.

Cependant, avant une possible utilisation de la mutation en sélection, des études complémentaires dans différentes lignées et testant son effet sur d'autres caractères seront nécessaires. L'objectif sera (1) d'évaluer la présence et d'estimer la fréquence de cette mutation au sein des différentes populations commerciales, et (2) d'évaluer si cette mutation peut affecter d'autres caractères que l'ELD (Epaisseur de Lard Dorsal).

Il n'est pas exclu que la mutation identifiée ne ségrége pas dans certaines races alors que cette région chromosomique contribue à la variabilité du caractère d'engraissement. Cette situation s'apparente à celle observée lors de l'identification de la mutation du gène GDF8 (gène codant pour la Myostatine) associé au phénotype d'hypertrophie musculaire chez les bovins (Grobet et al., 1997) ou du gène de prolificité BMP15 chez les ovins (Demars et al., 2013). Cette situation est due à une hétérogénéité allélique (différentes mutations du même gène aboutissent à un même phénotype). La stratégie sera alors de rechercher dans le gène responsable de la variabilité d'engraissement des animaux d'autres mutations fonctionnelles dans les différentes races ne portant pas la première mutation identifiée.

Ces analyses complémentaires permettront également d'évaluer si la mutation a un effet pléiotropique, et affecte simultanément la variabilité de différents caractères. Les travaux de thèse de Maxime Banville ont mis en évidence un effet de cette même région chromosomique pour un second caractère, influençant le nombre de tétines. Pour l'heure, nous ne savons pas si ces deux caractères sont régulés par une seule et même mutation ou si deux mutations différentes mais génétiquement liées affectent respectivement chacun de ces deux phénotypes.

Suivant la situation, les stratégies de sélection qui devront être mises en place devront être différentes, en fonction des populations et selon les objectifs de sélection respectifs pour les différents caractères affectés.

Un second intérêt plus fondamental sera d'évaluer l'intérêt de ces résultats pour les recherches sur l'obésité en génétique humaine. En effet, les modèles animaux jouent un rôle important pour une meilleure compréhension des mécanismes physiologiques et génétiques à l'origine des pathologies humaines. Les modèles murins les plus souvent utilisés ont une pertinence limitée comme modèle pour certaines maladies car les caractéristiques nutritionnelles, métaboliques et anatomiques de ces derniers sont éloignées de celles observées chez l'être humain. Au contraire, le porc présente de nombreuses similitudes avec l'homme et se révèle être un excellent modèle. On peut donc imaginer que lorsque le gène et la mutation causale seront identifiés, ces résultats puissent intéresser des cliniciens. En effet, dans l'état des connaissances actuelles, les 3 gènes candidats restants ne sont pas connus pour jouer un rôle dans ce type de maladies. Cela permettrait donc de s'intéresser à ces gènes comme gènes potentiellement candidats dans le métabolisme lipidique.

Pour conclure, en 2017 plus de 25 610 QTL ont été mis en évidence dans l'espèce porcine mais actuellement seul un très petit nombre de mutations ont été caractérisées.
Les ressources animales ont donc encore un large potentiel d'amélioration en déchiffrant les mécanismes moléculaires des caractères quantitatifs complexes.

Listes des références bibliographiques

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nature Methods *7*, 248–249.

Anderson, D.B., and Kauffman, R.G. (1973). Cellular and enzymatic changes in porcine adipose tissue during growth. J. Lipid Res. *14*, 160–168.

Anderson, S.I., Lopez-Corrales, N.L., Gorick, B., and Archibald, A.L. (2000). A large-fragment porcine genomic library resource in a BAC vector. Mamm. Genome *11*, 811–814.

Andersson, L., Haley, C.S., Ellegren, H., Knott, S.A., Johansson, M., Andersson, K., Andersson-Eklund, L., Edfors-Lilja, I., Fredholm, M., and Hansson, I. (1994). Genetic mapping of quantitative trait loci for growth and fatness in pigs. Science *263*, 1771–1774.

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

Archibald, A.L., Bolund, L., Churcher, C., Fredholm, M., Groenen, M.A.M., 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.

Asakawa, S., Abe, I., Kudoh, Y., Kishi, N., Wang, Y., Kubota, R., Kudoh, J., Kawasaki, K., Minoshima, S., and Shimizu, N. (1997). Human BAC library: construction and rapid screening. Gene *191*, 69–79.

Banville, M., Riquet, J., Bahon, D., Sourdioux, M., and Canario, L. (2015). Genetic parameters for litter size, piglet growth and sow's early growth and body composition in the Chinese-European line Tai Zumu. Journal of Animal Breeding and Genetics *132*, 328–337.

Berg, F., Stern, S., Andersson, K., Andersson, L., and Moller, M. (2006). Refined localization of the FAT1 quantitative trait locus on pig chromosome 4 by marker-assisted backcrossing. BMC Genetics *7*, 17.

Bidanel, J.P., Milan, D., Iannuccelli, N., Amigues, Y., Boscher, M.-Y., Bourgeois, F., Caritez, J.C., Gruand, J., Le Roy, P., Lagant, H., et al. (2000). Détection de locusa effets quantitatifs dans le croisement entre les races porcines Large White et Meishan. Résultats et perspectives. Journées de La Recherche Porcine En France *32*, 369–383.

Bidanel, J.-P., Milan, D., Iannuccelli, N., Amigues, Y., Boscher, M.-Y., Bourgeois, F., Caritez, J.-C., Gruand, J., Roy, P.L., Lagant, H., et al. (2001). Detection of quantitative trait loci for growth and fatness in pigs. Genetics Selection Evolution *33*, 289–310.

Boichard, D., Le Roy, P., Leveziel, H., and Elsen, J.-M. (1998). Utilisation des marqueurs moleculaires en genetique animale. INRA Prod. Anim *11*, 67–80.

Botstein, D., White, R.L., Skolnick, M., and Davis, R.W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. *32*, 314–331.

Cai, L., Taylor, J.F., Wing, R.A., Gallagher, D.S., Woo, S.S., and Davis, S.K. (1995). Construction and characterization of a bovine bacterial artificial chromosome library. Genomics *29*, 413–425.

Charlier, C., Coppieters, W., Farnir, F., Grobet, L., Leroy, P.L., Michaux, C., Mni, M., Schwers, A., Vanmanshoven, P., and Hanset, R. (1995). The mh gene causing double-muscling in cattle maps to bovine Chromosome 2. Mamm. Genome *6*, 788–792.

Cinti, S. (2005). The adipose organ. Prostaglandins Leukot. Essent. Fatty Acids *73*, 9–15.

Darvasi, A. (1998). Experimental strategies for the genetic dissection of complex traits in animal models. Nature Genetics *18*, 19–24.

Demars, J., Fabre, S., Sarry, J., Rossetti, R., Gilbert, H., Persani, L., Tosser-Klopp, G., Mulsant, P., Nowak, Z., Drobik, W., et al. (2013). Genome-Wide Association Studies Identify Two Novel BMP15 Mutations Responsible for an Atypical Hyperprolificacy Phenotype in Sheep. PLoS Genetics *9*, e1003482.

Fahrenkrug, S.C., Rohrer, G.A., Freking, B.A., Smith, T.P.L., Osoegawa, K., Shu, C.L., Catanese, J.J., and de Jong, P.J. (2001). A porcine BAC library with tenfold genome coverage: a resource for physical and genetic map integration. Mammalian Genome *12*, 472–474.

Fan, B., Du, Z.-Q., Gorbach, D.M., and Rothschild, M.F. (2010). Development and application of high-density SNP arrays in genomic studies of domestic animals. Asian-Australasian Journal of Animal Sciences *23*, 833–847.

Fève, B. (2005). Adipogenesis: cellular and molecular aspects. Best Pract. Res. Clin. Endocrinol. Metab. *19*, 483–499.

Filangi, O., Moreno, C., Gilbert, H., Legarra, A., Le Roy, P., and Elsen, J.M. (2010). QTLMap, a software for QTL detection in outbred populations. In Proceedings of the 9th World Congress on Genetics Applied to Livestock Production, p.

Foissac, S., Bardou, P., Moisan, A., Cros, M.-J., and Schiex, T. (2003). EUGENE'HOM: A generic similarity-based gene finder using multiple homologous sequences. Nucleic Acids Res. *31*, 3742–3745.

Fritz, S., Capitan, A., Djari, A., Rodriguez, S.C., Barbat, A., Baur, A., Grohs, C., Weiss, B., Boussaha, M., Esquerré, D., et al. (2013). Detection of Haplotypes Associated with Prenatal Death in Dairy Cattle and Identification of Deleterious Mutations in GART, SHBG and SLC37A2. PLoS ONE *8*, e65550.

Gellin, J., and Grosclaude, F. (1991). Analyse du génome des espèces d'élevage: projet d'établissement de la carte génétique du porc et des bovins. INRA Prod. Anim *4*, 97–105.

Georges, M. (2007). Mapping, Fine Mapping, and Molecular Dissection of Quantitative Trait Loci in Domestic Animals. Annual Review of Genomics and Human Genetics *8*, 131–162.

Gispert, M., Font i Furnols, M., Gil, M., Velarde, A., Diestre, A., Carrión, D., Sosnicki, A.A., and Plastow, G.S. (2007). Relationships between carcass quality parameters and genetic types. Meat Science *77*, 397–404.

González-Pérez, A., and López-Bigas, N. (2011). Improving the Assessment of the Outcome of Nonsynonymous SNVs with a Consensus Deleteriousness Score, Condel. The American Journal of Human Genetics *88*, 440–449.

Gotoh, T., Albrecht, E., Teuscher, F., Kawabata, K., Sakashita, K., Iwamoto, H., and Wegner, J. (2009). Differences in muscle and fat accretion in Japanese Black and European cattle. Meat Science *82*, 300–308.

Goureau, A. (1997). Contribution a l'etablissement de la carte genomique comparée entre l'Homme et le Porc (Sus Scrofa domestica), par "coloriage chromosomique". Université de Toulouse, Université Toulouse III-Paul Sabatier.

Goureau, A., Yerle, M., Schmitz, A., Riquet, J., Milan, D., Pinton, P., Frelat, G., and Gellin, J. (1996). Human and porcine correspondence of chromosome segments using bidirectional chromosome painting. Genomics *36*, 252–262.

Grobet, L., Martin, L.J., Poncelet, D., Pirottin, D., Brouwers, B., Riquet, J., Schoeberlein, A., Dunner, S., Ménissier, F., Massabanda, J., et al. (1997). A deletion in the bovine myostatin gene causes the double-muscled phenotype in cattle. Nat. Genet. *17*, 71–74.

Guillaume, F., Fritz, S., Boichard, D., and Druet, T. (2008). Estimation by simulation of the efficiency of the French marker-assisted selection program in dairy cattle *(Open Access publication)*. Genetics Selection Evolution *40*, 91–102.

Harismendy, O., Ng, P.C., Strausberg, R.L., Wang, X., Stockwell, T.B., Beeson, K.Y., Schork, N.J., Murray, S.S., Topol, E.J., Levy, S., et al. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biology *10*, R32.

Hospital, F., Chevalet, C., and Mulsant, P. (1992). Using markers in gene introgression breeding programs. Genetics *132*, 1199–1210.

Hu, Z.-L., Park, C.A., and Reecy, J.M. (2016). Developmental progress and current status of the Animal QTLdb. Nucleic Acids Research *44*, D827–D833.

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.

Illumina (2017). TruSeq Genotype Ne Reference Guide. 30.

Institut technique du porc (France) (2013). Mémento de l'éleveur de porc (Paris: IFIP-Institut du porc).

Jeon, J.T., Carlborg, O., Törnsten, A., Giuffra, E., Amarger, V., Chardon, P., Andersson-Eklund, L., Andersson, K., Hansson, I., Lundström, K., et al. (1999). A paternally expressed QTL affecting skeletal and cardiac muscle mass in pigs maps to the IGF2 locus. Nat. Genet. *21*, 157–158.

Jeon, J.T., Park, E.W., Jeon, H.J., Kim, T.H., Lee, K.T., and Cheong, I.C. (2003). A large-insert porcine library with sevenfold genome coverage: a tool for positional cloning of candidate genes for major quantitative traits. Mol. Cells *16*, 113–116.

Jiang, W., Zhao, X., Gabrieli, T., Lou, C., Ebenstein, Y., and Zhu, T.F. (2015). Cas9-Assisted Targeting of CHromosome segments CATCH enables one-step targeted cloning of large gene clusters. Nature Communications *6*, 8101.

Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature Protocols *4*, 1073–1081.

Lander, E.S., and Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics *121*, 185–199.

Laval, G., Iannuccelli, N., Legault, C., Milan, D., Groenen, M.A., Giuffra, E., Andersson, L., Nissen, P.H., Jørgensen, C.B., Beeckmann, P., et al. (2000). Genetic diversity of eleven European pig breeds. Genet. Sel. Evol. *32*, 187–203.

Le Dividich, J., Esnault, T., Lynch, B., Hoo-Paris, R., Castex, C., and Peiniau, J. (1991). Effect of colostral fat level on fat deposition and plasma metabolites in the newborn pig. J. Anim. Sci. *69*, 2480–2488.

Le Roy, P., and Elsen, J.M. (2000). Principes de l'utilisation des marqueurs génétiques pour la détection des gènes influençant les caractères quantitatifs. INRA Productions Animales 211–215.

Legault, C. (1978). Genetique et Reproduction chez le Porc. Journ Recher Porc 43–60.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Marklund, L., Nyström, P.E., Stern, S., Andersson-Eklund, L., and Andersson, L. (1999). Confirmed quantitative trait loci for fatness and growth on pig chromosome 4. Heredity (Edinb) *82 (Pt 2)*, 134–141.

Mason, I.L. (1988). A World Dictionary of Livestock Breeds, Types and Varieties (Oxford University Press).

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research *20*, 1297–1303.

McPherron, A.C., Lawler, A.M., and Lee, S.-J. (1997). Regulation of skeletal muscle mass in mice by a new TGF-p superfamily member. Nature *387*, 83–90.

Médigue, C., Bocs, S., Labarre, L., Mathé, C., and Vallenet, D. (2002). L'annotation in silico des séquences génomiques: Bio-informatique (1). Médecine/Sciences *18*, 237–250.

Mertes, F., ElSharawy, A., Sauer, S., van Helvoort, J.M.L.M., van der Zaag, P.J., Franke, A., Nilsson, M., Lehrach, H., and Brookes, A.J. (2011). Targeted enrichment of genomic DNA regions for next-generation sequencing. Briefings in Functional Genomics *10*, 374–386.

Milan, D., Woloszyn, N., Yerle, M., Le Roy, P., Bonnet, M., Riquet, J., Lahbib-Mansais, Y., Caritez, J.C., Robic, A., Sellier, P., et al. (1996). Accurate mapping of the "acid meat" RN gene on genetic and physical maps of pig chromosome 15. Mamm. Genome *7*, 47–51.

Milan, D., Bidanel, J.-P., Iannuccelli, N., Riquet, J., Amigues, Y., Gruand, J., Le Roy, P., Renard, C., and Chevalet, C. (2002). Detection of quantitative trait loci for carcass composition traits in pigs. Genetics Selection Evolution *34*, 705–728.

Milan, D., Demeure, O., Laval, G., Iannuccelli, N., Genet, C., Bonnet, M., Burgaud, G., Riquet, J., Gasnier, C., and Bidanel, J.-P. (2003). Identification de régions dugénome répondant à la sélection dans une lignée porcine sino-européenne: la Tai-zumu. Journ Recher Porc *35*, 309–316.

Montgomery, G.W., Sise, J.A., Greenwood, P.J., and Fleming, J.S. (1990). The Booroola F gene mutation in sheep is not located close to the FSH-beta gene. J. Mol. Endocrinol. *5*, 167–173.

Monziols, M., Bonneau, M., Mourot, J., Davenel, A., and Kouba, M. (2006). Les tissus adipeux intermusculaires présentent d'importantes particularités de développement et de composition en comparaison des tissus adipeux sous-cutanés chez le porc. Journées Recherche Porcine *38*, 61–66.

Mulsant, P., Lecerf, F., Fabre, S., Schibler, L., Monget, P., Lanneluc, I., Pisselet, C., Riquet, J., Monniaux, D., Callebaut, I., et al. (2001). Mutation in bone morphogenetic protein receptor-IB is associated with increased ovulation rate in Booroola Mérino ewes. Proc. Natl. Acad. Sci. U.S.A. *98*, 5104–5109.

Nezer, C., Moreau, L., Brouwers, B., Coppieters, W., Detilleux, J., Hanset, R., Karim, L., Kvasz, A., Leroy, P., and Georges, M. (1999). An imprinted QTL with major effect on muscle mass and fat deposition maps to the IGF2 locus in pigs. Nat. Genet. *21*, 155–156.

Nygard, A.-B., Jørgensen, C.B., Cirera, S., and Fredholm, M. (2007). Selection of reference genes for gene expression studies in pig tissues using SYBR green qPCR. BMC Molecular Biology *8*, 67.

Osoegawa, K., Woon, P.Y., Zhao, B., Frengen, E., Tateno, M., Catanese, J.J., and de Jong, P.J. (1998). An improved approach for construction of bacterial artificial chromosome libraries. Genomics *52*, 1–8.

Ott, J. (1999). Analysis of human genetic linkage (Baltimore: Johns Hopkins University Press).

Pfaffl, M.W. (2001). A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Res. *29*, e45.

Rafalski, A., and Morgante, M. (2004). Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. Trends in Genetics *20*, 103–111.

Riquet, J., Gilbert, H., Servin, B., Sanchez, M.-P., Iannuccelli, N., Billon, Y., Bidanel, J.-P., and Milan, D. (2011a). A locally congenic backcross design in pig: a new regional fine QTL mapping approach miming congenic strains used in mouse. BMC Genetics *12*, 6.

Riquet, J., Mercat, M., Iannuccelli, N., Servin, B., Pailhoux, E., and Larzul, C. Recherche de causes génétiques des anomalies congénitales majeures chez le porc. 7.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nature Biotechnology *29*, 24–26.

Rogel-Gaillard, C., Bourgeaux, N., Billault, A., Vaiman, M., and Chardon, P. (1999). Construction of a swine BAC library: application to the characterization and mapping of porcine type C endoviral elements. Cytogenet. Cell Genet. *85*, 205–211.

Rohrer, G.A., and Keele, J.W. (1998). Identification of quantitative trait loci affecting carcass composition in swine: I. Fat deposition traits. J. Anim. Sci. *76*, 2247–2254.

Sanchez, M.-P., Riquet, J., Iannuccelli, N., Gogue, J., Billon, Y., Demeure, O., Caritez, J.-C., Burgaud, G., Feve, K., Pery, C., et al. (2005). Programme de cartographie fine de QTL. Journ Recher Porc *35*, 65–72.

Sanchez, M.-P., Riquet, J., Iannuccelli, N., Gogue, J., Billon, Y., Demeure, O., Caritez, J.-C., Burgaud, G., Feve, K., Bonnet, M., et al. (2006). Effects of quantitative trait loci on chromosomes 1, 2, 4, and 7 on growth, carcass, and meat quality traits in backcross Meishan x Large White pigs. Journal of Animal Science *84*, 526–537.

Sanchez, M.-P., Tribout, T., Iannuccelli, N., Bouffaud, M., Servin, B., Dehais, P., Muller, N., Mercat, M.-J., Estelle, J., Bidanel, J.-P., et al. (2012). Cartographie fine de régions QTL à l'aide de la puce Porcine SNP60 pour l'ingestion, la croissance, la composition de la carcasse et la qualité de la viande en race Large White. In 44. Journées de La Recherche Porcine. 2012-02-072012-02-08, Paris, FRA, (IFIP-Institut du Porc; INRA), p.

Schibler, L., Vaiman, D., Oustry, A., Guinec, N., Dangy-Caye, A.L., Billault, A., and Cribiu, E.P. (1998). Construction and extensive characterization of a goat bacterial artificial chromosome library with threefold genome coverage. Mamm. Genome *9*, 119–124.

Schibler, L., Vaiman, and Cribiu (2000). Origine du polymorphisme de l'ADN. INRA Prod. Anim.

Schwob, S., Riquet, J., Bellec, T., Kernaleguen, L., Tribout, T., and Bidanel, J.-P. (2009). Mise en place d'un programme de sélection assistée par marqueurs dans la population sino-européenne Duochan. Journées Rech. Porcine *41*, 29–30.

Servin, B., Faraut, T., Iannuccelli, N., Zelenika, D., and Milan, D. (2012). High-resolution autosomal radiation hybrid maps of the pig genome and their contribution to the genome sequence assembly. BMC Genomics *13*, 585.

Suzuki, K., Asakawa, S., Iida, M., Shimanuki, S., Fujishima, N., Hiraiwa, H., Murakami, Y., Shimizu, N., and Yasue, H. (2000). Construction and evaluation of a porcine bacterial artificial chromosome library. Anim. Genet. *31*, 8–12.

Thierry-Mieg, D., and Thierry-Mieg, J. (2006). AceView: a comprehensive cDNA-supported gene and transcripts. Genome Biology *7*, S12.

Tortereau, F., Servin, B., Frantz, L., Megens, H.-J., Milan, D., Rohrer, G., Wiedmann, R., Beever, J., Archibald, A.L., Schook, L.B., et al. (2012). A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. BMC Genomics *13*, 586.

Trayhurn, P., Temple, N.J., and Van Aerde, J. (1989). Evidence from immunoblotting studies on uncoupling protein that brown adipose tissue is not present in the domestic pig. Can. J. Physiol. Pharmacol. *67*, 1480–1485.

Tribout, T. (2013). Intérêt de la sélection génomique dans les programmes de sélection porcins: cas d'une lignée mâle de grande taille.

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.

Vignal, A., Milan, D., SanCristobal, M., and Eggen, A. (2002). A review on SNP and other types of molecular markers and their use in animal genetics. Genetics Selection Evolution *34*, 275–305.

Visscher, H., Brinkhuis, H., Dilcher, D.L., Elsik, W.C., Eshet, Y., Looy, C.V., Rampino, M.R., and Traverse, A. (1996). The terminal Paleozoic fungal event: evidence of terrestrial ecosystem destabilization and collapse. Proc. Natl. Acad. Sci. U.S.A. *93*, 2155–2158.

Yerle, M., Pinton, P., Robic, A., Alfonso, A., Palvadeau, Y., Delcros, C., Hawken, R., Alexander, L., Beattie, C., Schook, L., et al. (1998). Construction of a whole-genome radiation hybrid panel for high-resolution gene mapping in pigs. Cytogenet. Cell Genet. *82*, 182–188.

Yerle, M., Pinton, P., Delcros, C., Arnal, N., Milan, D., and Robic, A. (2002). Generation and characterization of a 12,000-rad radiation hybrid panel for fine mapping in pig. Cytogenet. Genome Res. *97*, 219–228.

Zhu, C., Gore, M., Buckler, E.S., and Yu, J. (2008). Status and Prospects of Association Mapping in Plants. The Plant Genome Journal *1*, 5.

Résumé

CARTOGRAPHIE FINE ET CARACTERISATION D'UN QTL LOCALISE SUR LE CHROMOSOME 1 PORCIN INFLUENCANT L'ADIPOSITE DES ANIMAUX.

La mise en évidence de gènes affectant la croissance et l'engraissement des animaux présente un fort intérêt pour la filière porcine. Les gènes influençant la variabilité de ce type de caractère sont appelés QTL, pour Quantitative Trait Loci. Au début des années 1990, l'INRA a initié un programme de détection de QTL affectant les principaux caractères d'intérêt économique (croissance, engraissement, reproduction…).

Un QTL influençant l'adiposité des animaux a été mis en évidence sur le chromosome 1 porcin à partir d'un croisement F2 entre des verrats Large White et des Truies Meishan. Pour ce QTL, l'allèle Meishan est associé à une plus forte adiposité (Q), et les allèles Large White sont dominants sur les allèles de la race Meishan. Par conséquent, des individus hétérozygotes présenteront des phénotypes proches des individus homozygotes Large White.

L'objectif de ce travail a consisté à affiner la cartographie de ce locus d'intérêt, pour cela plusieurs approches complémentaires ont été utilisées :

- Un dispositif animal de type back-cross a été mis en place afin de réduire significativement l'intervalle de localisation du QTL.
- Une étude transcriptomique a été réalisée à partir de deux dispositifs expérimentaux pour mener une première approche fonctionnelle. Des gènes différentiellement exprimés ont été recherchés dans 3 tissus (longe, foie et tissu adipeux) en comparant les génotypes des animaux à 3 stades (10, 30 et 110Kg).
- Enfin, un séquençage haut-débit a été entrepris pour répertorier l'ensemble des variants présents dans la région du QTL. Dans un premier temps, un séquençage tout génome a été réalisé pour les deux individus recombinants définissant les bornes basse et haute de l'intervalle du QTL. Puis dans un second temps, le séquençage uniquement de la région, sur un plus grand nombre d'individus pour lesquels un testage sur descendance avait été effectué, afin de s'assurer de leur statut au QTL.

Aujourd'hui même si la mutation candidate n'a pas pu être encore identifiée, ces différentes approches nous ont permis de réduire de façon considérable la région du QTL. En effet, nous sommes passés d'une région de 1,2 Mb à 280 Kb, comprenant encore quelques centaines de variants. Les premières analyses bio-informatiques laissent présager que nous allons pouvoir réduire fortement ce nombre et envisager des approches fonctionnelles plus ciblées.

Mots clés : *QTL, chromosome 1, Large white, Meishan, épaisseur de lard dorsal, cartographie génétique, back-cross, séquencage NGS.*