

**LA COMBINAISON DE CARTOGRAPHIE DE SIGNATURES DE SELECTION,  
DE RE-SEQUENCAGE DE GENOMES ET D'ANALYSES D'EXPRESSION REVELE  
PARK2 et JAG2 COMME NOUVEAUX GENES CANDIDATS  
REGULANT L'ADIPOSITE**

**Roux Pierre-François<sup>\*,§,†</sup>, Boitard Simon<sup>‡,\*\*</sup>, Blum Yuna<sup>§§</sup>, Parks Brian<sup>§§</sup>, Montagner Alexandra<sup>††</sup>,  
Moussel Etienne<sup>‡‡</sup>, Djari Anis<sup>\*\*\*</sup>, Esquerré Diane<sup>§§§</sup>, Désert Colette<sup>\*,§,†</sup>, Boutin Morgane<sup>\*,§,†</sup>, Leroux  
Sophie<sup>†††</sup>, Lecerf Frédéric<sup>\*,§,†</sup>, Le Bihan-Duval Elisabeth<sup>‡‡‡</sup>, Klopp Christophe<sup>\*\*\*</sup>, Servin Bertrand<sup>†††</sup>,  
Pitel Frédérique<sup>†††</sup>, Duclos Michel Jean<sup>‡‡‡</sup>, Guillou Hervé<sup>††</sup>, Lusi Aldons J.<sup>§§,\*\*\*\*</sup>, Demeure Olivier<sup>\*,§,†</sup>,  
Lagarrigue Sandrine<sup>\*,§,†,1</sup>**

<sup>\*</sup>INRA, UMR1348 Pegase, Saint-Gilles, F-35590. <sup>§</sup>Agrocampus Ouest, UMR1348 Pegase, Rennes, F-35000. <sup>†</sup> Université Européenne de Bretagne. <sup>‡</sup>UMR 7205 OSEB (MNHN – CNRS – EPHE – UPMC), Paris, F-75005. <sup>\*\*</sup>INRA/AgroParisTech, UMR1313 GABI, Jouy-en-Josas, F-78352. <sup>§§</sup>Department of Medicine, UCLA, Los Angeles, USA. <sup>††</sup>UMR1331, INP, UPS, TOXALIM, INRA, F-31000 Toulouse. <sup>‡‡</sup>INSERM, UMR 1048 IMPC F-31432 Toulouse. <sup>\*\*</sup>INRA, SIGENAE, Castanet-Tolosan, F-31326. <sup>§§§</sup>INRA, Plateforme GENOTOUL, Castanet-Tolosan, F-31326. <sup>†††</sup>UMR INRA, INPT ENSAT, INPT ENVT-GENPHYSE, Castanet-Tolosan, F-31326. <sup>‡‡‡</sup>INRA, UR83 Recherches Avicoles, Nouzilly, F-37380. <sup>\*\*\*\*</sup>Department of Human Genetics, UCLA, Los Angeles, USA.

[sandrine.lagarrigue@agrocampus-ouest.fr](mailto:sandrine.lagarrigue@agrocampus-ouest.fr)

## RESUME

Très peu de gènes responsables de la variation de caractères quantitatifs ont été identifiés par les méthodes de cartographie de QTLs (Quantitative Trait Loci) en raison de la grande taille de ces régions. La plupart des gènes identifiés jusqu'alors l'ont été en raison d'un lien fonctionnel déjà connu avec le phénotype ciblé. Cette étude propose de combiner des approches de détection de QTLs et de signatures de sélection, d'annotation de SNPs, et d'analyses d'expression pour faciliter l'identification des gènes causaux sous-jacents aux régions QTLs.

L'étude a été menée sur des lignées expérimentales de poulets sélectionnées de façon divergente pour un seul caractère complexe, le poids de gras abdominal, et pour lesquelles plusieurs QTLs ont déjà été cartographiés. En utilisant une nouvelle approche statistique exploitant des densités très élevées de SNPs et basée sur les fréquences haplotypiques, nous avons identifié 129 signatures de sélection. La plupart des QTLs co-localisent avec au moins une signature de sélection ; cette co-localisation réduit alors considérablement la taille de ces régions QTLs. Certaines de ces régions ne contiennent plus qu'un seul gène, faisant de ces derniers de sérieux candidats sans aucun *a priori* fonctionnel. Nous nous sommes ensuite focalisés sur deux de ces régions QTLs particulièrement réduites en taille. L'absence de SNP non-synonyme dans les régions codantes suggère fortement l'existence de mutations causales agissant en *cis* sur l'expression des gènes associés à ces régions. L'analyse de l'expression de ces gènes dans les lignées divergentes de poulets mais aussi dans des modèles murins génétiquement contrastés pour l'adiposité confirme leur liaison avec le phénotype d'intérêt. Cette étude montre pour la première fois l'intérêt de combiner des approches basées sur la détection de traces de sélection, l'annotation de SNPs et l'analyse d'expression dans des lignées divergentes sélectionnées pour un caractère spécifique et met en évidence deux nouveaux gènes *JAG2* et *PARK2* comme de potentiels régulateurs clés, négatif et positif respectivement, de l'adiposité chez le poulet et la souris.

## ABSTRACT

### Combined QTL and selective sweep mapping with genome re-sequencing and expression revealed *PARK2* and *JAG2* as new candidate genes for adiposity regulation

Very few causal genes have been identified by quantitative trait loci (QTLs) mapping because of the large size of QTLs, and most of them were identified thanks to functional links already known with the targeted phenotype. Here we propose to combine selection signature detection, SNP annotation, and expression analyses to identify causal genes underlying QTLs. As a model, we chose experimental chicken lines divergently selected for only one trait, the abdominal fat weight, in which several QTLs were previously mapped. Using a new haplotype-based statistics exploiting the very high SNP density generated through whole genome re-sequencing, we found 129 significant selective sweeps. Most of the QTLs co-localized with at least one sweep, which markedly narrowed candidate region size. Some of those sweeps contained only one gene, therefore making them strong positional causal candidates with no presupposed function. We then focused on two of these QTLs/sweeps. The absence of non-synonymous SNP in their coding regions strongly suggests the existence of causal mutations acting in *cis* on their expression, confirmed for one of both. Additional expression analyses on those two genes in the divergent chicken lines and in genetically diverse mice contrasted for adiposity confirm their link with this phenotype. This study shows for the first time the interest of using selective sweeps combined with SNPs and expression analyses in divergent lines selected for a specific trait for identifying causative genes and highlights two genes, *JAG2* and *PARK2*, as new potential negative and positive key regulators of adiposity in chicken and mice.

## INTRODUCTION

Au cours des 30 dernières années, des centaines de QTLs affectant différents caractères pour de nombreuses espèces ont été cartographiés. Dans ce contexte, alors que les analyses d'association sur l'ensemble du génome (GWAS) se développent comme méthode de référence, la plupart des QTLs déjà identifiés dans les espèces d'élevage l'ont été à partir d'analyses de liaison, le plus souvent dans des croisements de type F<sub>2</sub>. C'est en particulier vrai chez le poulet où le coût de production d'un individu est peu élevé et l'intervalle de génération court, ce qui a permis de créer des lignées divergentes pour certains caractères d'intérêt, lignées ensuite croisées pour maximiser l'hétérozygotie aux QTLs dans les protocoles expérimentaux. A ce jour, 4714 QTLs sont référencés dans la Mouse Genome Database alors que 8006, 8935 et 3926 le sont respectivement pour le bovin, le porc et le poulet dans la base de données animal QTL database. Le principal défaut de ces QTLs cartographiés par analyse de liaison est leur taille, le plus souvent de plusieurs mégabases (Mbs). Ces intervalles de localisation contiennent des centaines de gènes, compliquant fortement l'identification des gènes causaux. Lorsque l'on étudie ces larges régions génomiques, il est donc tentant de ne considérer que les gènes candidats fonctionnels. C'est pourquoi la plupart des gènes causaux identifiés jusqu'ici étaient déjà connus pour leur rôle dans le caractère associé (Le Bihan-Duval *et al.* 2011), ce qui limite l'intérêt de ces recherches. Dans cette étude, nous proposons d'identifier les signatures de sélection entre deux lignées de poulets divergentes pour le poids de tissu adipeux abdominal en utilisant des millions de SNPs obtenus par le re-séquençage de génome d'animaux de ces lignées. Pour utiliser au mieux cette densité de marqueurs, nous avons utilisé la méthode statistique hapFLK (Fariello *et al.* 2013) qui mesure la différenciation génétique entre échantillons sur la base des fréquences des haplotypes et non des fréquences des allèles, lui permettant ainsi de prendre en compte la structure de corrélation entre les SNPs. Un autre avantage important à utiliser les données de DNA-seq est la disponibilité quasi exhaustive des polymorphismes (indels et SNPs) caractérisant les individus d'intérêt. Parmi les gènes candidats positionnels identifiés dans des QTL/traces de sélection, la disponibilité de données de type DNA-seq permet de trier les gènes en deux catégories comme indiqué en Figure 1 : ceux avec un SNP ou un indel impactant la protéine mature (renforçant le statut potentiellement causal du gène et de la mutation) et ceux avec des polymorphismes qui pourraient agir en *cis* sur leur expression (requérant alors des analyses d'expression tissulaire des gènes pour statuer sur leur statut de candidat). Nous proposons dans cette étude de combiner ces différentes approches génétiques et expressionnelles

pour faciliter l'identification de gènes causaux pour l'adiposité.

## 1. MATERIELS ET METHODES

### Animaux

Les deux lignées expérimentales de poulets de chair ont été sélectionnées de façon divergente pendant 7 générations sur le pourcentage de gras abdominal à 9 semaines d'âge, en gardant constant le poids vif (Leclercq *et al.* 1980). Les deux lignées ont été nommées lignées grasse (LG) et maigre (LM). Vingt animaux issus de la 35<sup>ème</sup> génération ont été séquencés : 7 LM, 4 LG et 9 F<sub>1</sub> (LG x LM, animaux non apparentés) incluant les 5 mâles F<sub>1</sub> du dispositif expérimental utilisé pour la détection des QTLs.

### Re-séquençage du génome de 20 animaux

Le re-séquençage de l'ADN génomique a été effectué par fragments de 2 x 100 bp sur l'appareil HiSeq 2000 (Illumina). Les séquences ont été alignées contre le génome de référence de poule WASHUC2.1 en utilisant le logiciel BWA v0.7.0 puis filtrées sur la qualité de séquences et les duplicats de PCR (SAMtools v0.1.19). Les SNPs & indels ont été alors identifiés en utilisant GATK UnifiedGenotyper.

### Caractérisation fonctionnelle des SNPs

L'annotation de l'impact fonctionnel des SNPs sur les protéines a été réalisée à l'aide du logiciel Variant Effect Predictor du site bioinformatique « Ensembl ».

### Détection de traces de sélection avec 9,4 M de SNPs

La détection des régions soumises à sélection entre les LG et LM se fait à l'aide de l'outil statistique hapFLK de Fariello *et al.* (2013). Dans un premier temps, une distribution neutre (sous dérive génétique) des statistiques de tests est estimée sur la base des valeurs observées à l'échelle du génome, la majorité des loci du génome étant neutres. Les quelques régions sous sélection ont à l'inverse des valeurs extrêmes par rapport à cette distribution. Les régions sous sélection considérées comme significatives sont celles ayant une q-value inférieure à 0,1 (soit une tolérance de 10% de fausses traces de sélection).

### Analyses d'expression différentielle

L'ARN total des foies et tissus adipeux de 24 animaux pré-pubères (9 semaines d'âge) des LG et LM ont été extraits avec du TRIzol. Le niveau d'expression génique a été obtenu par RT-PCR quantitative et normalisé avec le gène *GAPDH*. La différence de niveau d'expression d'un gène entre les génotypes est testée par un test t de Student.

### Expression dit ‘allèle spécifique’

Cette étude consiste à tester si une position hétérozygote donnée d’un gène génère les deux transcrits correspondants en quantités équivalentes. Pour cela, huit ADN génomiques et ARN hépatiques des poulets F<sub>1</sub> (LM x LG) ont été analysés à deux positions : chr3:46.581.638 pb et chr3:46.581.695 bp, en utilisant un séquenceur Qiagen PyroMark Q24. Seuls 5 animaux F<sub>1</sub> étaient hétérozygotes aux deux positions et ont été conservés pour les analyses. La significativité du déséquilibre d’expression entre « allèles » est déterminée par un test non-paramétrique de Mann-Whitney en comparant le ratio allélique observé à celui attendu (égal à 1).

## 2. RESULTATS ET DISCUSSION

Le premier objectif de cette étude était d’identifier des traces de sélection en utilisant une forte densité de SNPs identifiées et génotypées par le re-séquençage de génomes entiers d’animaux issus de lignées divergentes pour le poids de tissu adipeux abdominal.

### Détection des SNPs dans les génomes des LG et LM

Le séquençage de l’ADN génomique de 20 animaux a produit environ 400 Gb de données brutes et a permis d’identifier 9,4M de SNP et 1,1M d’indels (insertions et délétions) dans les deux lignées de poulets de chair Grasse et Maigre. On observe en moyenne 2,7M de SNPs par poulet, soit  $2.6 \pm 0.5$  SNPs/kb, ce qui est en accord avec des études antérieures (Wong *et al.* 2004). Comme attendu, les SNPs ont été principalement identifiés dans les régions intergénomiques (48,7%) et introniques (41,8%) alors que les régions régulatrices et codantes ne représentent respectivement que 8,4% et 1,2%. Parmi les SNP des régions codantes, 25%, 0,45% et 15% ont pour conséquences respectives un changement d’acide aminé, le gain ou la perte d’un codon stop ou d’un site d’épissage.

### Identification de 129 empreintes de sélection

En utilisant la statistique de test hapFLK (Fariello *et al.* 2013) qui est particulièrement adaptée à l’utilisation de millions de SNPs pour étudier des populations avec de petits effectifs, 129 traces de sélection ont été mises en évidence. Ces régions sont également distribuées sur le génome et ont une taille moyenne de 97,5 kb, contenant en moyenne 838 SNPs et 2,11 gènes. Un enrichissement significatif en gènes impliqués dans le métabolisme des lipides a été observé en étudiant les 52 régions ne contenant qu’un seul gène ( $p < 10^{-14}$ , test de  $\chi^2$ ). Ce résultat renforce l’idée que les régions détectées sont des régions affectées par la pression de sélection appliquée lors de la sélection divergente des deux lignées. Nous avons

ensuite superposé les régions soumises à sélection avec les régions QTL affectant l’engraissement abdominal détectées en 2006 pour déterminer les traces de sélection les plus pertinentes.

### Confrontation des régions QTL et des traces de sélection

Nous avons précédemment observé 6 QTLs pour l’engraissement abdominal en utilisant un dispositif F<sub>2</sub> issu du croisement des lignées LM et LG et composé de 5 familles de mâles F<sub>1</sub> et de 585 F<sub>2</sub> (Lagarrigue *et al.* 2006). Quatre de ces 6 QTLs co-localisent avec au moins une trace de sélection (parfois jusqu’à cinq traces) contenant de 1 à 11 gènes (Figure 2 : voir les QTLs notés AF3.I, AF3.II, AF5 et AF7 respectivement sur les chromosomes 3, 3, 5 et 7. Ainsi, l’utilisation des traces de sélection a permis de réduire considérablement les intervalles de localisation des QTLs, passant en moyenne de 12 Mb à 100 kb. Certaines régions ne contenant plus qu’un gène, ce dernier est dès lors un candidat positionnel très fort. Cette importante réduction de taille représente donc une réelle avancée pour l’identification des gènes causaux sous-jacents à des QTL. Nous avons alors focalisé nos efforts sur les régions AF5 et AF7 contenant peu de gènes candidats (respectivement JAG2 pour AF5 et MLLT4 et PARK2 pour AF7. Pour des raisons de clarté, nous présenterons ici les résultats concernant uniquement le QTL AF7 sur le chromosome 7.

### Focus sur le QTL AF3.II contenant

#### MLLT4 et PARK2

Ce QTL inclut 3 traces de sélection (Figure 2), l’une d’entre elle contenant certainement une mutation causale. Ces trois traces de sélection contiennent respectivement aucun gène, une partie de *MLLT4* et une partie de *PARK2*. Pour la première trace de sélection, la confrontation avec les données RNA-seq disponibles dans la base de données Ensembl a permis de confirmer l’absence de gène dans cette région. Concernant, les gènes *MLLT4* et *PARK2*, aucun indel ou mutation non synonyme pouvant avoir un impact sur la protéine n’a été observé, suggérant donc une mutation causale agissant en *cis* sur l’expression d’au moins un de ces deux gènes. Le foie et le tissu adipeux (TA) étant des acteurs majeurs du métabolisme des lipides, ces deux tissus ont été choisis pour les études d’expression de ces deux gènes dans les lignées LG et LM. Au préalable nous avons vérifié que ces deux gènes étaient exprimés dans ces tissus. A la différence de *MLLT4*, un différentiel d’expression significatif a été obtenu pour *PARK2* dans les deux tissus, avec une expression plus importante dans la LG (Figure 3A). Une expression différentielle peut être due à deux mécanismes : un mécanisme de type « trans », c’est à dire faisant intervenir une autre protéine (par exemple un facteur

de transcription) codée par un gène ailleurs dans le génome, cette protéine amont étant plus ou moins fonctionnelle entre les deux lignées. Dans ce cas, le gène différentiellement exprimé entre lignées n'est qu'une cible d'un régulateur plus amont : il n'est donc pas le gène régulateur recherché. Le second mécanisme est de type « cis », il fait référence à l'existence d'un variant dans les régions régulatrices du gène différentiellement exprimé, ce variant étant ainsi directement responsable de la variation d'expression observée ; ce peut être par exemple un SNP dans le promoteur du gène. Dans ce cas, le gène différentiellement exprimé entre lignées est lui-même responsable de son différentiel d'expression et représente alors un gène régulateur très pertinent pour le caractère d'intérêt. Une régulation de type « cis » (situation qui nous intéresse) peut être observée par des technologies permettant de tester si l'expression est allèle spécifique. Nous montrons que *PARK2* présente une expression allèle spécifique dans le foie pour les deux marqueurs étudiés (Figure 3B), avec une expression là aussi plus importante pour l'allèle spécifique de la LG. Ces résultats indiquent donc l'existence d'un variant agissant en *cis* sur l'expression de *PARK2* dans le foie, et renforce le statut de gène causal de ce gène pour la régulation de l'adiposité. Ces résultats sont par ailleurs cohérents avec ceux obtenus sur les modèles murins, dans lesquels on observe un doublement de l'expression de *PARK2* dans le foie et le TA de souris obèses comparé à des souris sauvages (Figure 3C). Ces corrélations positives entre l'expression de *PARK2* avec

l'adiposité corporelle, observées chez le poulet et la souris indique que *PARK2* est un régulateur positif de l'adiposité. L'haplotype fixé lors de la sélection divergente l'ayant été dans la lignée grasse (Figure 2B), cet haplotype correspondant à un allèle « gain de fonction ».

## CONCLUSION

Cette étude montre l'intérêt d'utiliser les traces de sélection dans le contexte de cartographie de QTLs à partir de lignées sélectionnées de façon divergente sur un caractère d'intérêt. Nous montrons que la co-localisation entre QTLs et traces de sélection réduit considérablement la taille de ces régions QTLs. Certaines de ces régions ne contiennent plus qu'un seul gène, faisant de ces derniers de sérieux candidats sans aucun *a priori* fonctionnel. Par ailleurs nous montrons l'intérêt de combiner des approches de génétique avec des approches expressionnelles pour renforcer le statut « causal » de ces gènes. Nous avons ainsi mis en évidence *JAG2* (*données non montrées*) et *PARK2* comme deux nouveaux potentiels régulateurs, respectivement négatif et positif, de l'adiposité chez le poulet mais aussi la souris (*données non montrées*).

**Keywords :** signature de sélection, re-séquencage de génomes, annotation des SNP, expression, adiposité

**Session ciblée :** Génétique

F1rnaSEQ-2012) et l'agence nationale de la recherche (Fatinteger 2012-2015). La bourse doctorale de Pierre-François Roux est co-financée par l'INRA et la Région Bretagne.

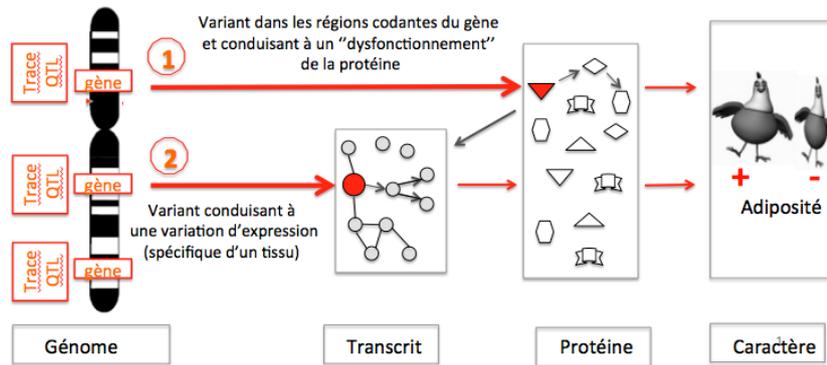
## REMERCIEMENTS

Nous remercions le personnel de l'UE1295 PEAT, Nouzilly, France pour l'élevage des poulets. Ces travaux ont été financés par l'INRA (ChickSeq-2010,

## REFERENCES BIBLIOGRAPHIQUES

1. Fariello, M. I., S. Boitard, H. Naya, M. SanCristobal, and B. Servin. 2013 Detecting Signatures of Selection Through Haplotype Differentiation Among Hierarchically Structured Populations. *Genetics* 193:929–941.
2. Lagarrigue, S., F. Pitel, W. Carre, B. Abasht, P. Le Roy, et al. 2006 Mapping quantitative trait loci affecting fatness and breast muscle weight in meat-type chicken lines divergently selected on abdominal fatness. *Genet. Sel. Evol.* 38:85–97.
3. Leclercq, B., J. C. Blum, and J. P. Boyer. 1980 Selecting broilers for low or high abdominal fat. *Br. Poult. Sci.* 21:107–113.
4. Le Bihan-Duval, E., J. Nadaf, C. Berri, F. Pitel, B. Graulet, et al. 2011 Detection of a Cis eQTL Controlling BMCO1 Gene Expression Leads to the Identification of a QTG for Chicken Breast Meat Color. *PLoS ONE* 6:e14825.
5. Wong, G. K.-S., B. Liu, J. Wang, Y. Zhang, X. Yang, et al. 2004 A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. *Nature* 432:717–722.

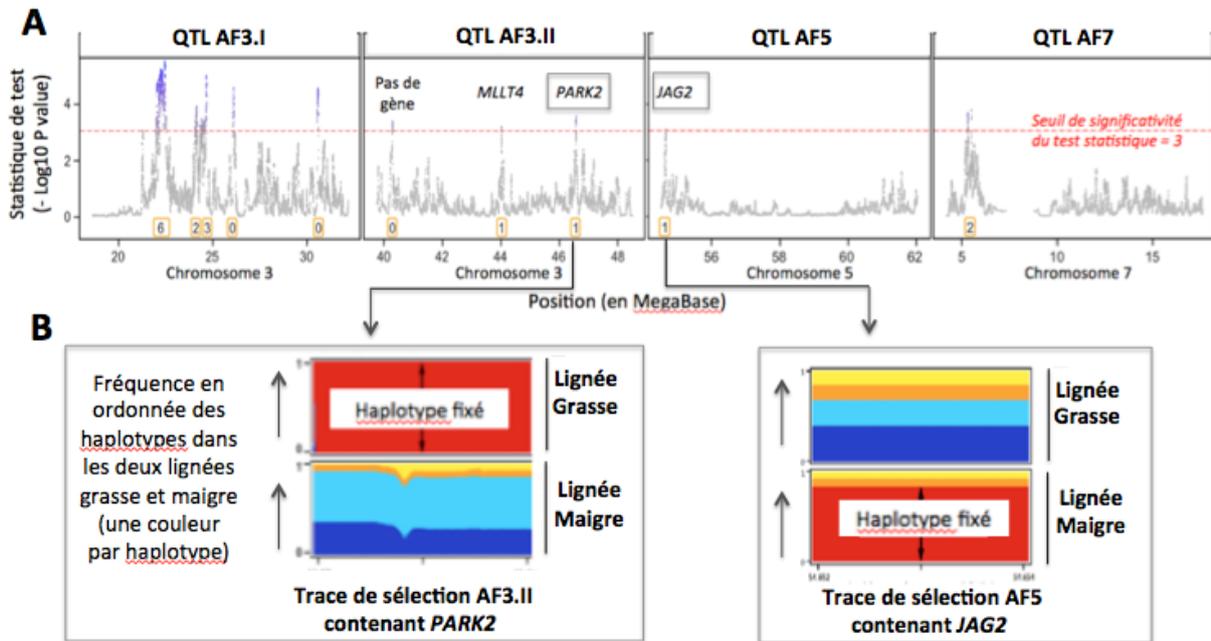
**Figure 1 :** Deux grands types de variations dans le génome pouvant conduire à une variation d'un caractère d'intérêt.



Les gènes causaux sous jacents aux QTL ou traces de sélection responsables des variations du caractère d'intérêt sont indiqués dans un encadré rouge. 1. Variants (SNP ou indel) dans les régions codantes du gène "causal" impactant la protéine mature associée indiquée par un triangle rouge.

2. Variant dans les régions régulatrices du gène "causal" conduisant à une variation de l'expression du gène, visible au niveau de ses transcrits indiqués par un cercle rouge.

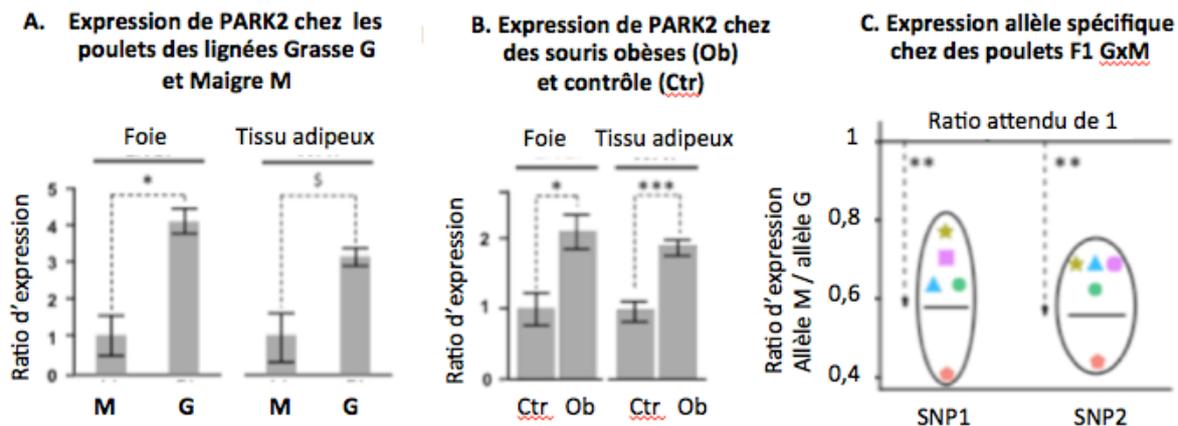
**Figure 2 :** Traces de sélection colocalisant avec 4 QTLs d'adiposité corporelle.



**A)** Les positions génomiques des 4 QTLs sont en abscisses des 4 encadrés ; le test statistique de détection d'une trace de sélection est en ordonnée ; les traces significatives sont en violet (test statistique > 3). Les nombres en encadré sont le nombre de gènes présents dans la trace de sélection correspondante.

**B)** Fréquences des haplotypes (indiqués par des couleurs) dans chaque lignée pour les deux traces de sélection étudiées en détail par l'équipe, soit AF3.II et AF5 contenant respectivement PARK2 et JAG2.

**Figure 3 :** Différentes observations en faveur du statut de *PARK2* comme gène régulateur positif de l'adiposité.



**A.** Poulets maigre (M) & gras (G). **B.** Souris mâles et adultes obèses (Ob) ou contrôle (CTR). **C.** Deux variants (SNP1 et SNP2) dans le gène PARK2 (exon3) montrant que le nombre de copie d'ARNm PARK2 chez des croisés F1 GxM est plus important pour la copie chromosomique spécifique de la lignée Grasse que celle spécifique de la lignée Maigre (ratio d'expression M/G < 1). Ce ratio est en cohérence avec le différentiel d'expression observé en A entre des poulets de lignée Grasse versus lignée Maigre.

\* : pvalue < 0,05; \*\* : pvalue < 0,01 ; \*\*\* : pvalue < 0,001 \$ : pvalue < 10%.