Contrôle génétique

de l’adiposité corporelle

Etude de l’espèce Gallus Gallus

**M2 BIS - UE NGS**

**William Dartois – Vivien Dupont**

# Introduction

L’espèce Gallus Gallus est aujourd’hui, et depuis longtemps, très étudiée en génétique et génomique de par son intérêt agronomique. Cette espèce est principalement divisée en deux groupes : les poulets de chair, ayant une masse musculaire importante et un taux de gras faible, et les poulets de ponte. Chez ces derniers, la capacité à stocker le gras pour pouvoir l’utiliser ultérieurement, notamment dans des conditions environnementales plus strictes, est un mécanisme clé dans l’allongement de la période de ponte. Dans les deux cas, le métabolisme des lipides est mis en jeu.

Dans la littérature, certaines études indiquent que des régions génomiques situées sur les chromosomes 8, 17 et 24 porteraient des gènes partiellement impliqués dans l’adiposité corporelle. Ainsi, dans l’optique de maîtriser au mieux cette caractéristique, nous avons analysé le chromosome 17, via des données RNAseq extraites du foie et du tissu adipeux, de huit espèces Gallus Gallus (quatre lignées grasse (FL) et quatre lignées mince (LL)).

Avec cette étude, nous cherchons à déterminer un gène ou variant candidat, impliqué dans le mécanisme de variation de l’adiposité corporelle.

# Analyse et interprétation

## Données étudiées

Tableau 1: Caractéristiques des données utilisées

|  |  |
| --- | --- |
|  | |
| **Région génomique** | Chromosome 17 |
| **Séquençage** | - RNA sequencing  - Sélection polyA  - Paired-end |
| **Taille des reads** | 100 pb (x2) |
| **Lignées** | - FL : F5 ; F7 ; F10 ; F13  - LL : F6 ; F8 ; F9 ; F11 |
| **Fichiers à disposition** | - 8 fastq (4 FL ; 4 LL)  - Génome de référence (« .fa »)  - Annotation du génome de référence (« .gtf ») |
| **Taille de la région chromosomique** | 10,956,400 pb |
| **Nombre de gènes** | 368 |
| **Nombre d’exons** | 6135 |
| **SNP** | 509 |

Commentaires :

→ Expliquer notamment comment on arrive aux fichier VEP et VariantTotable.

→ Expliquer ce que contiennent ces fichiers

## Sélection des gènes/variants d’intérêts

Une fois que l’ensemble du traitement des données est effectué, depuis l’alignement des fichiers fastq sur le génome de référence, jusqu’à l’obtention du fichier VCF, il s’agit à présent de prioriser les variants ou gènes d’intérêt. Pour cela, deux approches sont possibles :

* Une approche classique de sélection des variants potentiellement intéressants, suite aux prédictions réalisées par L’outils VEP (Variant Effect Predictor) d’Ensembl.
* Une approche de sélection sans a priori, avec recherche des variants dont les fréquences des génotypes sont contrastées entre les lignées FL et LL. Cette approche sera complétée d’une analyse GWAS pour confirmer les résultats.

### Sélection par contraste de fréquences

Suite à l’utilisation des outils HaplotypeCaller, CombineGVCFs et GenotypeGVCFs, nous avons obtenus et sélectionné, dans un premier temps, les variants correspondants à des SNP ou à des SNPcluster. Ces résultats, comme expliqué précédemment, sont présent dans un fichier au format VCF. Grâce à la suite *GATK,* ce fichier VCF a été transformé en un fichier « .table », que l’on peut directement étudier avec le langage R, sous Rstudio.

Quelques modifications du fichier ont été effectuées avant de calculer les fréquences des génotypes :

* Remplacement les « | » par des « / » comme séparateur des deux allèles d’un génotype. En effet, le format de représentation des génotypes est le suivant : « allèle1/allèle2 ». Cependant, il est possible que le slash soit remplacé par un pipe dans le cas d’un phasage avec un autre SNP.
* Ajout d’une variable indiquant sur le variant est phasé ou non.
* Suppression de tous les variants dont l’allèle de référence, ou alternatif, est différent de « A », « C », « G » ou « T ».
* Codage des génotypes :

- ALT/ALT devient 1/1

- ALT/REF et REF/ALT deviennent respectivement 1/0 et 0/1

- REF/REF devient 0/0

* Calcul des fréquences des génotypes (0/0, 0/1, 1/1), pour chaque variants.

Suite à ces modifications, et dans l’optique de trouver des variants dont les fréquences des génotypes sont contrastées entre les deux lignées étudiées, une sélection des variants 100 % « 0/0 » dans une lignée, et 100 % « 0/1 » ou « 1/1 » dans l’autre, ont été sélectionné. En effet, il y a seulement quatre individus par lignée, nous ne voulons donc pas retrouver l’allèle alternatif dans la lignée qui n’est pas représentatif de la variation, et toujours le retrouver dans l’autre lignée, pour qu’il y ait un réel contraste.

Avec ces contraintes, 18 variants potentiels ont été obtenu. Pour affiner notre recherche, il a été décidé d’éliminer les variants avec une faible profondeur de séquençage, car ils peuvent être sensibles aux erreurs de séquençage. Un seuil à 80 lectures est alors choisi. En dessous de ce seuil, chez certains individus le SNP a une profondeur de séquençage inférieure à 10, et c’est trop faible pour avoir la certitude que ce n’est pas dû à une erreur de séquençage.

13 variants candidats ont alors été obtenu. Dans un premier temps, à l’aide du logiciel IGV, nous avons été regarder l’emplacement du variant (exons, introns…) sur le génome de référence, et vérifié si celui-ci était phasé ou non avec une autre variation (Table 2). Ensuite, ...

Tableau 2: Caractéristiques des variants candidats sélectionnés selon les fréquences contrastées des génotypes entre les deux lignées

| **Lignée** | **Position** | **Variants** | **Phasage** | **Mutation** | **Gènes** | **Candidat** |
| --- | --- | --- | --- | --- | --- | --- |
| LL | 5432425 | C → T | NON | Silencieuse | FPGS | OUI |
| LL | 5434744 | G → A | NON | 3’ downstream | CDK9 | OUI |
| LL | 5434899 | G → A | NON | 3’ downstream | CDK9 | OUI |
| LL | 5434950 | G → A | NON | 3’ downstream | CDK9 | OUI |
| LL | 5435648 | G → A | NON | 3’ downstream | CDK9 | OUI |
| LL | 5444037 | G → A | NON | Exon non référencé, entre deux gènes : 5’upstream SH2D3C (silencieuse, avant codon start) ou CDK9 (Arginine remplacée par une Glycine) | SH2D3C ou CDK9 | OUI potentiel |
| LL | 5444069 | C → T | NON | Exon non référencé, entre deux gènes : 5’upstream SH2D3C (silencieuse, avant codon start) ou CDK9 (Arginine remplacée par une Glycine) | SH2D3C ou CDK9 | OUI potentiel |
| LL | 5513019 | T → C | NON | Silencieuse | SLC27A4 | OUI |
| LL | 5513895 | A → G | NON | Silencieuse | SLC27A4 | OUI |
| LL | 5514499 | C → T | OUI | Aspartate remplacée par une Asparagine | SLC27A4 | OUI |
| LF | 5528011 | T → C | OUI | 3’ downstream | ENSGALT0000007872 | OUI |
| LF | 5528254 | T → C | NON | Exon non référencé, entre deux gènes : ENSGALT0000007872 et ENSGALT00000086495 | ENSGALT0000007872 ou ENSGALT00000086495 | OUI Potentiel |
| LF | 5528341 | G → A | NON | Exon non référencé, entre deux gènes : ENSGALT0000007872 et ENSGALT00000086495 | ENSGALT0000007872 ou ENSGALT00000086495 | OUI Potentiel |

Il y de cela quelque temps, le variant dans le gène FPGS n’aurait pas été candidat. Cependant, aujourd’hui on sait qu’une mutation silencieuse peut tout de même avoir un impact, notamment sur le repliement de la protéine, du fait de la modification de la vitesse de traduction. Pour cette raison, et dans l’optique de ne pas passer à côté d’un potentiel variant impliqué dans la régulation de l’adiposité corporelle, il a été décidé de le conserver comme potentiel candidat.

### Sélection par GWAS

L’une des techniques statistiques les plus utilisées pour l’analyse de données génomiques est le GWAS (Genome-wide association study) qui consiste à étudier pour chaque variation génétique chez les individus la corrélation de ce variant avec un trait phénotypique, en l’occurrence ici, l’adiposité corporelle.

Afin de mener à bien le GWAS, il a fallu organiser nos données en un tableau de données adapté. Pour ce faire on organise la table de la manière suivante : Chaque ligne représente désormais un SNP à une certaine position pour un individu. Par exemple :

Tableau 3: Base de données en ligne

|  |  |  |  |
| --- | --- | --- | --- |
| Individu | GT | Grosseur | Position |
| 5 | 0/0 | FL | 5432425 |
| 6 | 1/1 | LL | 5432425 |

Il y a donc désormais 8 lignes par SNP. Ce tableau de données permet dorénavant de réaliser les régressions nécessaires au GWAS. On remarque que notre variable d’intérêt (la Grosseur) est binaire (soit FL soit LL), il faut donc ajuster une régression logistique pour tenir compte de la nature de cette variable. Le facteur explicatif (le fait d’être 0/0, 0/1 ou 1/1) est une variable catégorielle, on choisit la modalité 0/0 pour référence afin de permettre l’identifiabilité du modèle. De plus on exclut du GWAS les SNP où les 8 poules étaient toutes 0/1 ou 1/1 puisque les SNP associés n’expliquent évidemment pas les différences de grosseur entre les poules.

Pour tous les 177 SNP restants on réalise une régression de la façon suivante :

**Régressions logistiques pour tous les SNP :**

**Arguments** : {« FL », « LL »}

= {« 0/0 », « 0/1 », « 1/1 »}, le fait d’être ref/ref, ref/alt ou alt/alt

**Pour** j l’ensemble des positions des SNP:

**Faire la régression logistique :**

**Tester :**  « » vs « »

**Stocker :** la p-valeur du test

**Fin pour**

**Retourner :** Le vecteur de p-valeur de tous les tests

On n’applique pas de correction sur le vecteur de p-valeur final puisque le faible nombre d’individu entraine des p-valeurs pas assez faible. Le GWAS n’est évidemment pas adapté pour si peu d’individu (On considère qu’il en faut au moins quelques milliers) mais dans un but de comparaison entre les différentes méthodes de sélection de SNP les résultats du GWAS sont intéressants.

Au final, le GWAS sélectionne, parmi les p-valeurs inférieures à 0.004 (2.4 en -log10), tous les SNP choisis par la méthode des fréquences (en rouge dans le graphique) ainsi que ceux non choisit car pas d’assez bonne qualité (DP<80).

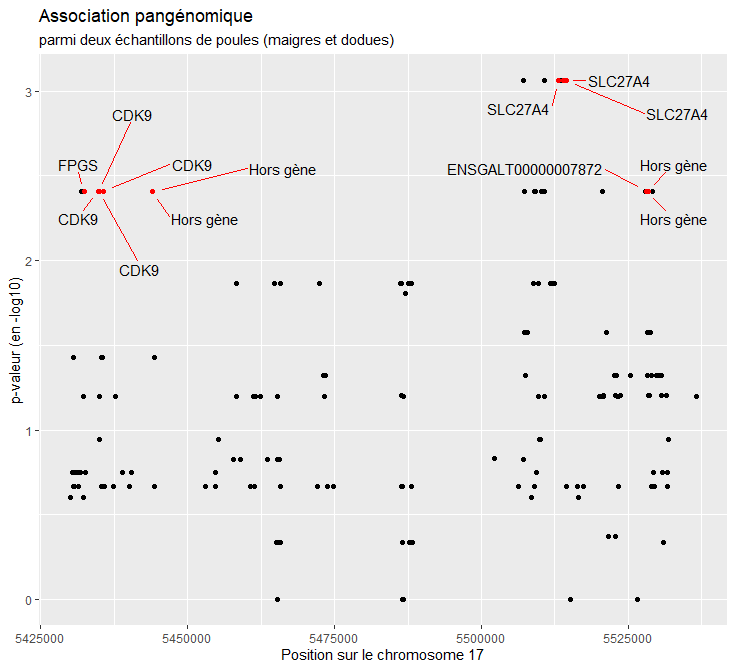


Figure 1: Résultats du GWAS

### Sélection après analyse VEP

Cette approche est la plus intuitive et la plus simple, car l’ensemble des prédications sur les variants est réalisé en amont par l’outil VEP.

Dans un premier temps, seules les variants avec un impact fort ont été conservé, il en restait dix. Ensuite, pour affiner notre recherche, les variants dont les génotypes ne sont pas contrastés entre les deux lignées, ont été écartés (Tableau 3).

Tableau 4: Fréquence des génotypes des variants à fort impacte

| **Position** | **LF**  **0/0** | **LF**  **0/1** | **LF**  **1/1** | **LL**  **0/0** | **LL**  **0/1** | **LL**  **1/1** | **Contraste** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 5458993 | 0,5 | 0,25 | 0,25 | 0 | 0,75 | 0,25 | NON |
| 5472646 | 0,75 | 0 | 0,25 | 0,25 | 0,5 | 0,25 | NON |
| 5488885 | 0 | 0 | 0 | 0 | 0 | 0,5 | NON |
| 5501417 | 0 | 0 | 0,25 | 0 | 0 | 0 | NON |
| 5501491 | 0 | 0 | 0,25 | 0 | 0 | 0,25 | NON |
| 5511517 | 0,5 | 0 | 0,25 | 0,5 | 0 | 0,25 | NON |
| 5514455 | 0 | 0,25 | 0 | 0 | 0 | 0,25 | NON |
| 5515091 | ? | ? | ? | ? | ? | ? | NON |
| 5515128 | 1 | 0 | 0 | 0 | 1 | 0 | OUI |
| 5532695 | 0 | 0 | 0,5 | 0 | 0 | 0,5 | NON |

Seul un variant est conservé suite à l’analyse du contraste des fréquences génotypique (Table 3). Il a ensuite été déterminé que ce variant pouvait être un candidat pour être responsable de la variation du métabolisme de l’adiposité corporelle (Table 4), grâce notamment à l’analyse des conséquences de la variation sous IGV.

Tableau 5: Caractéristiques du variant en position 5515128

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Lignée** | **Position** | **Variants** | **Phasage** | **Mutation** | **Gènes** | **Candidat** |
| LL | 5515128 | G → A | OUI | Arginine remplacée par une cystéine | SLC27A4 | OUI |

Ce variant est phasé avec un autre SNP, dans le même exon. Néanmoins, ce que l’on peut voir sous IGV (Illustration 1), est que, malgré le second SNP, les lignées LF codent toujours pour une arginine alors que les lignées LL codent pour une arginine ou une cystéine (car la variation G → A ségrége toujours avec la variation T → G dans le même exon).

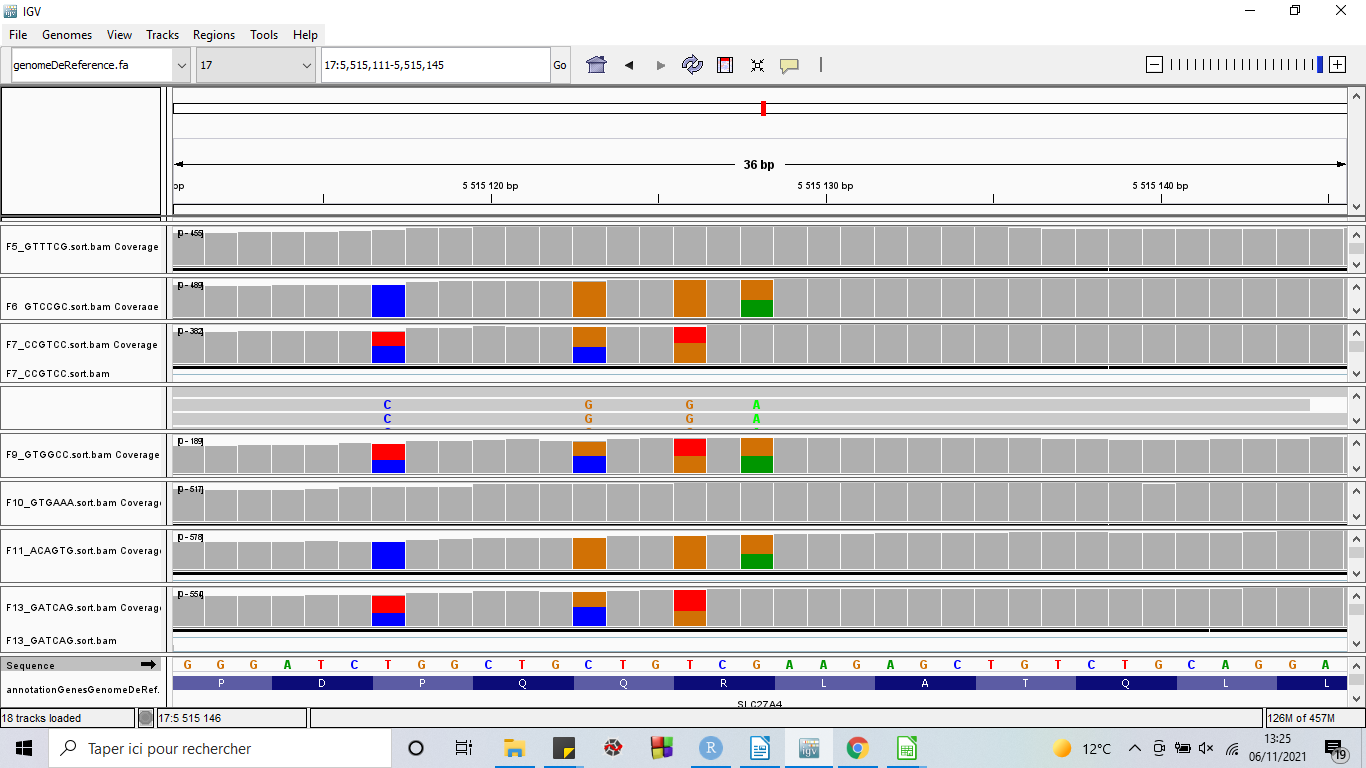


Figure 2: capture d’écran IGV de la position 5515128 sous IGV

## Fonctions associées aux variants candidats

ENSGALT0000007872 :

TRUB2 :

- synthèse d’ARNm pseudourine

- Régulation positive des mécanismes de traduction mitochondriaux

- Traitement de l’ARN

CDK9 :

- Serine/Thréonine kinase (régulation positive de l’ubiquitination histone H2B, de la phosphorylation d’histones, régulation de la réparation de l’ADN, régulation de la différenciation des cellules musculaires, régulation de la transcription, régulation négative de la polyadénylation).

SH3D2C :

- Régulation positive de la phosphorylation peptidyl-serine

- Transduction de signal médiée par GTPase

SLC27A4 :

FPGS :

- biosynthèse des composés contenant de l’acide folique

- Synthèse du glutamate

- Synthèse du tetrahydrofolylpolyglutamate

Nécessaire pour la synthèse de composé organique

2.4\_ Validation du rôle des variants candidats dans la régulation de l’adiposité corporelle