

UTILISATION DES NGS POUR LA CARACTERISATION GENETIQUE DE LIGNEES DE POULETS DE CHAIR ET POULES DE PONTE

P.F. Roux^{1,2}, F. Lecerf^{1,2}, A. Djari³, D. Esquerre⁴, S. Marthey⁵, T. Zerjal⁶, E. Le Bihan-Duval⁷, C. Klopp³, M. Moroldo⁵, J. Estelle⁶, C. Désert^{1,2}, B. Bed'hom⁶, M. Tixier-Boichard⁶, O. Demeure^{1,2}, S. Lagarrigue^{1,2}

¹ INRA, UMR1348 PEGASE, Rennes, France

² Agrocampus Ouest, UMR1348 PEGASE, Rennes, France

³ INRA, SIGENAE, Castanet Tolosan, France

⁴ INRA, Plateforme GENOTOUL, Castanet Tolosan, France

⁵ INRA, UMR1313 GABI, CRB GADIE, Jouy-en-Josas, France

⁶ INRA, UMR1313 GABI, Jouy-en-Josas, France

⁷ INRA, UR83 Recherches Avicoles, Nouzilly, France

sandrine.lagarrigue@agrocampus-ouest.fr

RÉSUMÉ

Les nouvelles technologies de séquençage appelées « NGS » offrent aujourd'hui la possibilité d'analyser les génomes dans leur intégralité. Dans ce contexte nouveau, cette étude vise à caractériser à l'échelle du génome les polymorphismes de lignées expérimentales INRA de poulets de chair et poules de ponte, divergentes respectivement pour le poids de gras abdominal (G/M) ou l'efficacité alimentaire résiduelle (R+/R-). Les génomes d'une quinzaine d'individus par lignée ont été séquencés soit en pool (R+/R-) soit individuellement (G/M) avec une profondeur de 19X sur un HiSeq 2000. Dans un premier temps, nous présentons quelques statistiques concernant les 11 Millions de SNP identifiés sur l'ensemble de ces lignées. Par ailleurs, 386 177 et 510 577 SNP discriminant respectivement les individus R+ des R- et les individus G des M ont été identifiés dont 1,5% sont localisés dans des gènes et sont susceptibles d'avoir un impact fonctionnel. L'étude fine des fonctions des gènes associés conduit à de nouvelles pistes d'étude sur les causes de la divergence entre lignées. Enfin, ces approches nous ont permis d'identifier une mutation *a priori* fonctionnelle dans un gène candidat pour un QTL affectant le gras abdominal.

ABSTRACT

New sequencing technologies called « NGS » allow the analysis of whole genomes. In this new context, our study aims at characterizing at the genome scale polymorphisms within INRA broilers and layers experimental lines divergent for abdominal fatness (G/M) and feed efficiency (R+/R-), respectively. Genomes of about 15 animals per line were sequenced either in pools (R+/R-) or individually (G/M), using a HiSeq 2000 and with a depth of 19X. We first present some summary statistics about the 11 million SNPs identified in the 4 lines. In addition, 386,177 SNP discriminating R+ from R- and 510,577 SNPs discriminating G from M were identified. About 1.5% of these SNP are localized in genes, possibly with a functional modification. A fine study of genes functions leads to new hypotheses on how the divergence between lines occurred. Finally, these approaches allowed us to identify one mutation that might be functional in a candidate gene for a QTL affecting abdominal fatness.

INTRODUCTION

Les nouvelles technologies de séquençage appelées « NGS » offrent aujourd'hui la possibilité d'analyser les génomes dans leur intégralité. De par ces technologies, les coûts de séquençage ont diminué d'un facteur 2000 ces dernières années. Ces technologies permettent notamment d'identifier les polymorphismes les plus fréquents d'une lignée ou d'une sous population dans les espèces déjà séquencées. En effet, la diminution des coûts de séquençage permet d'envisager le re-séquençage de quelques dizaines d'individus à un prix raisonnable voire de beaucoup plus d'individus si la région à séquencer est ciblée. L'identification des polymorphismes de lignées présente plusieurs intérêts : on peut citer la recherche de polymorphismes causaux dans des gènes candidats de régions QTL d'intérêt, la recherche de polymorphismes spécifiques de lignées, la recherche de signature de sélection ou encore l'identification de polymorphismes non synonymes pouvant avoir un impact sur des processus biologiques. Peu d'études utilisant de telles approches ont été publiées à ce jour. Notons cependant l'étude de Rubin et al., (2010) reportant plus de 7 millions de SNP observés au sein de 9 populations de poules. Ces auteurs ont identifié la mutation causale déterminant un QTL dans des lignées divergentes grâce au re-séquençage d'échantillons mélangés.

Dans ce contexte nouveau, notre étude vise à caractériser à l'échelle du génome (environ 1 milliard de bases (1Gb) chez la poule) les polymorphismes de quatre lignées expérimentales de l'INRA. Ces quatre lignées sont issues de deux programmes de sélection divergente pour des caractères d'intérêt pour lesquels des programmes de détection de QTL ont été réalisés. Les lignées concernées sont des lignées de poulets de chair et poules de ponte divergentes respectivement pour le poids de gras abdominal (G/M) et l'efficacité alimentaire résiduelle (R+/R-). Concernant les polymorphismes, seuls les SNP (Single Nucleotide Polymorphism) sont caractérisés, laissant les INDEL (Insertions-délétions) ou encore les CNV (Copy Number Variation) à des analyses ultérieures.

1. MATERIELS ET METHODES

1.1. Procédures expérimentales

Concernant les lignées G et M, 16 individus (9 F1 GxM, 3 F0 de la lignée M et 4 BC) ont été re-séquencés. Notons que 8 individus F0 des deux lignées sont en cours de séquençage. Concernant les lignées R+/R-, deux pools correspondant à 7 coqs R+ et 7 coqs R- non apparentés ont été re-séquencés. Les ADN génomiques ont été extraits à partir de sang. Les banques de séquençage ont été préparées selon la technique Illumina. Les fragments de 100 pb en « paired ends » ainsi générés ont été séquencés sur un séquenceur HiSeq 2000 (Illumina) sur 18 lignes au total (soit 20 Gb par individu).

1.2. Identification des SNP

Les lectures issues du séquençage d'ADN des 18 échantillons ont été alignées sur la version Galgal3 issue de l'assemblage 2.1 du génome de poule réalisé par le *Washington University Genome Sequencing Center* et datant de Mai 2006. Les variants de type SNP ont été alors identifiés avec entre autres comme critère une qualité des bases ≥ 30 . Pour ce faire, la suite logicielle Samtools [Li et al., 2009] a été utilisée dans un premier temps donnant des résultats très hétérogènes d'un individu à l'autre. Dans un second temps la suite logicielle GATK [McKenna et al., 2010] a été utilisée, permettant un meilleur contrôle des alignements et une meilleure calibration de chaque base le long des lectures.

1.3. Identification des SNP spécifiques de lignées

Une base de données MySQL répertoriant les SNP identifiés et génotypes associés a été créée afin de faciliter les analyses ultérieures. Un certain nombre de statistiques a été généré (nombre de SNP par individu, nombre total, fréquences des allèles,...). La sélection des SNP spécifiques de lignées a été réalisée selon les critères suivants : concernant les lignées G et M, les 9 individus F1 devaient être tous hétérozygotes et les 3 individus F0 devaient être homozygotes. Concernant les lignées R+ et R-, des SNP avec un allèle différent fixé dans chacun des pools ont été recherchés.

1.4. Caractérisation fonctionnelle des SNP spécifiques de lignées

Les annotations des SNP ont été réalisées via l'application « *Variant Effect Predictor* » d'Ensembl. Cette application permet de localiser un SNP dans les différentes régions d'un gène : en amont ou en aval de la zone transcrite (zone

upstream), dans les zones codantes (traduites en protéine) ou encore dans les zones transcrites mais non traduites (UTR pour UnTranslated Region). Pour les SNP des régions codantes, l'impact fonctionnel du variant est également indiqué (introduction d'un codon stop, affectation des sites d'épissage, ...).

1.5. Annotation des gènes ayant un SNP spécifique de lignées

Les termes fonctionnels associés à chaque gène d'intérêt ont été collectés depuis les bases de données Gene Ontology. L'orthologue humain a été utilisé afin de collecter le plus grand nombre de termes. Pour chaque liste de gènes d'intérêt, les termes dits surreprésentés (i.e. non présents par le hasard) ont été recherchés. L'ensemble de ces tâches a été réalisé à l'aide de l'outil DAVID (Huang et al., 2009).

2. RESULTATS ET DISCUSSION

Identification des SNP

Toute nouvelle technologie demande à la communauté scientifique un délai de quelques années pour créer des outils d'analyse stables et efficaces. En 2011, une première identification des SNP a été faite avec les outils les plus utilisés du moment à savoir le programme Mpileup de la suite logicielle Samtools. Nous avons identifié un nombre de SNP très hétérogène entre individus, allant de 800 000 à 3M avec une moyenne de 1,3 M. Par ailleurs moins de 10% des positions présentaient un génotype renseigné sur plus de 50% des individus. Ces premières analyses indiquaient une procédure d'analyse inappropriée à nos données de séquences. En 2012, a été réalisée une seconde analyse fondée sur les outils de la suite logicielle GATK qui introduit des étapes de réalignement et recalibration permettant pour certains jeux de données de diminuer les biais liés au séquençage et/ou à la préparation des échantillons. En moyenne 2 152 128 et 3 051 799 SNP ont été respectivement identifiés par individu G/M et R+/R- soit une variation respective sur tout le génome de $3,05 \pm 0,28$ et $2,15 \pm 0,1$ SNP/kb pour les deux populations. Ces résultats sont cohérents avec la littérature (International Chicken Polymorphism Map Consortium 2004). A l'échelle des deux populations analysées, 11 198 092 SNP ont ainsi été identifiés dont 9 556 735 (85,3%) présentaient un génotype renseigné sur 100% des individus analysés. Parmi ces positions, 5 385 843 et 8 793 699 sont

respectivement polymorphes chez les R+/R- et les G/M dont 4 437 610 en commun (82,4% des R+/R-). Ces résultats sont conformes à l'histoire de ces deux populations, les G/M ayant été créés dans les années 70 à partir de différentes lignées de chair alors que les R+/R- sont des poules de ponte correspondant à une seule lignée, la *Rhode Island Red*.

Identification des SNP spécifiques de lignées R+ versus R- ou G versus M

Parmi les positions polymorphes des deux populations, 510 577 SNP discriminent la LG de la LM (soit 5,6%) et 386 177 SNP discriminent la R- de la R+ (soit 7,1%); 15990 sont communs aux deux fonds génétiques R-/R+ et LG/LM. Ces allèles fixés sont le fruit de la sélection opérée sur les caractères poids de gras abdominal et efficacité alimentaire mais également de la dérive génétique. Cette dernière doit être importante compte tenu de l'effectif réduit de nos populations expérimentales et de la durée de la sélection divergente. Il serait donc intéressant de rechercher les régions sous sélection positive afin d'estimer la part de la dérive génétique et de la sélection dans la fixation des allèles. Cette approche de recherche de signature de sélection est en cours dans les lignées R+ et R-. Les résultats préliminaires montrent qu'un nombre fini de régions (quelques dizaines) apparaissent sous forte sélection positive dans ces lignées divergentes.

Caractérisation fonctionnelle de ces SNP et des gènes associés

Parmi les SNP spécifiques de lignées, un certain nombre est responsable des différences phénotypiques entre lignées divergentes, et ce d'autant plus qu'ils se trouvent dans des gènes. Parmi les SNP discriminant les G des M et les R+ des R-, on en dénombre respectivement dans les deux populations 2229 (0,4%) et 1486 (0,3%) dans les régions codantes et 5054 (1%) et 3561 (0,9%) dans les régions non codantes (UTR et région promotrice proximale définie à 500 bases) (**Tableau 1**). Les premiers que l'on nommera SNP « codants » sont susceptibles de conduire à un changement de structure de la protéine alors que les seconds, nommés « régulateurs », peuvent affecter la transcription ou la stabilité des transcrits et donc au final la quantité de protéine. Comme attendu, ces SNP « codants » et « régulateurs » sont rares (moins de 1,5%). La grande majorité (98%) des allèles différenciellement fixés entre lignées divergentes

sont donc dans des régions introniques ou inter-géniques et sont probablement en grande partie neutres ; ces résultats suggèrent un impact fort de la dérive génétique connue pour être importante dans des petites populations.

Les lignées G/M et R+/R- divergent pour le poids de gras abdominal et l'efficacité alimentaire résiduelle ; ce qui laisse penser qu'elles portent un certain nombre de polymorphismes ayant des conséquences fonctionnelles pour des gènes impliqués dans le métabolisme cellulaire et plus précisément dans le métabolisme énergétique et notamment lipidique. Afin de tester cette hypothèse, nous avons réalisé une analyse des fonctions des gènes associés aux SNP « codants » et « régulateurs » ayant *a priori* un effet sur la structure ou la quantité des protéines associées. Nous avons identifiés 3327 (cas des G/M) et 1319 (cas des R+/R-) gènes associés à ces SNP. Les termes fonctionnels sur-représentés dans ces listes de gènes sont présentés dans le **tableau 2**. Parmi les 7 termes fonctionnels reportés pour chaque population, 2 à 3 termes sont associés à des métabolismes énergétiques. De façon intéressante d'autres termes par exemple associés pour les R+/R- à la réponse immunitaire sont trouvés. De telles observations peuvent être probablement expliquées par une combinaison de différents facteurs. La dérive génétique joue probablement un rôle en fixant de façon aléatoire des SNP dans des régions fonctionnelles. Ce peut être également le fruit de la sélection qui sur un métabolisme aussi complexe que le métabolisme de l'énergie peut opérer sur d'autres processus biologiques en lien avec ce métabolisme. Ainsi, ces termes sont autant de processus biologiques pouvant être impliqués dans les divergences phénotypiques des lignées G versus M et R+ versus R-. Ces deux populations divergent pour le poids de gras abdominal. Par ailleurs les 234 SNP (171 gènes) spécifiques de lignées et communs aux deux populations sont associés à différents termes fonctionnels dont deux en lien avec le métabolisme des lipides.

Intérêt pour l'identification de mutations causales dans des régions QTL

Une région QTL affectant le poids et le pourcentage de tissu adipeux abdominal chez les lignées G et M a été identifiée sur le chromosome 19 grâce à un dispositif F2 (Demeure et al. 2013). Cette région porte le gène acetyl-CoA carboxylase A (ACACA) qui est largement décrit pour son importance dans le

métabolisme des lipides. Une étude eQTL montre que l'expression de ce gène n'est pas contrôlée par cette région suggérant une très probable mutation dans les régions codantes ou d'épissage de ce gène. L'analyse fine de l'ensemble des SNP dans le gène ACACA a permis d'identifier un polymorphisme non synonyme dont le génotype chez les F1 est concordant avec le génotype de ces derniers au QTL. Cette mutation est par ailleurs très conservée entre espèces et affecte le domaine carboxyl transférase ayant une activité catalytique déterminante dans le fonctionnement de cette protéine (**Figure 1**). Notons que ce variant n'est pas spécifique d'une des lignées conformément aux résultats de détection de QTL où il a été observé la non fixation des allèles dans les lignées fondatrices. Ce faisceau d'observations concordantes suggère un impact de ce polymorphisme sur la fonction de la protéine ACACA. Des études expérimentales complémentaires sont cependant à mener pour valider cette hypothèse. Cet exemple montre tout l'intérêt de dresser le catalogue des SNP d'une lignée d'intérêt pour laquelle des analyses de détection de QTL ont été mises en place.

A l'inverse, il est possible d'analyser immédiatement la présence ou non de polymorphismes déjà identifiés par ailleurs comme affectant un caractère d'intérêt. Nous avons par exemple recherché les deux variations affectant la transcription du gène BCMO1 responsable de variation de la couleur de la viande (Le Bihan-Duval et al 2011). Ces mutations sont absentes dans les deux populations G/M et R+/R- au moins dans les échantillons analysés.

CONCLUSION

Sur le plan technologique, cette étude montre l'importance de suivre l'évolution des outils bioinformatiques mais aussi de les comparer afin d'utiliser les plus appropriés au jeu de données analysé. On attend prochainement d'autres évolutions en terme de séquençage (fragments séquencés plus longs) qui demanderont alors de nouveaux outils et/ou paramétrages des outils actuels pour le pré-traitement de ces séquences. Sur le plan biologique, cette étude a permis de dresser un premier catalogue des SNP des lignées de poulets de chair G/M et d'une lignée de ponte R+/R- ainsi qu'une liste des variants spécifiques des lignées divergentes G versus M et R+ versus R-. Une étude fine des fonctions de

ces variants localisés dans les régions géniques nous donnent de nouvelles hypothèses de travail et montre l'avantage des approches NGS pour détecter sans *a priori* des processus biologiques impliqués dans des divergences phénotypiques. Par ailleurs, comme l'illustre l'exemple du gène ACACA, ces approches de re-séquençage devraient faciliter l'identification de mutations causales candidates sous-jacentes à des QTL. Cependant l'annotation fonctionnelle des

variants dans les régions régulatrices reste un frein dans ce type d'études. Enfin un certain nombre d'observations suggère un rôle important de la dérive génétique dans nos deux populations G/M et R+/R- d'effectif réduit. Des études seront donc à mener pour distinguer les effets de la dérive génétique de ceux de la sélection dans ces populations.

Mots clés : NGS séquençage poule SNP

REMERCIEMENTS : Ce travail a été réalisé grâce au soutien de l'INRA (programme Chick-seq). P.-F. Roux bénéficie d'une bourse de thèse Région Bretagne/INRA. Merci à Alan Jinq, étudiant d'Agrocampus Ouest pour sa participation à la création de la base de données de SNP.

REFERENCES BIBLIOGRAPHIQUES

- Rubin CJ, Zody MC, Eriksson J, Meadows JR, Sherwood E, Webster MT, Jiang L, Ingman M, Sharpe T, Ka S, Hallböök F, Besnier F, Carlborg O, Bed'hom B, Tixier-Boichard M, Jensen P, Siegel P, Lindblad-Toh K, Andersson L. 2010. Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature*, (7288):587-91.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R., 2009. 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, (16):2078-9
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, (9):1297-303
- Huang da, W., Sherman, B.T., Lempicki, R.A., 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* (4) 44–57.
- International Chicken Polymorphism Map Consortium, 2004. A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. *Nature* (432) 717–722.
- Demeure et al. 2013. JRA-JRPFG2013/000186
- Le Bihan-Duval E, Nadaf J, Berri C, Pitel F, Graulet B, Godet E, Leroux SY, Demeure O, Lagarrigue S, Duby C, Cogburn LA, Beaumont CM, Duclos MJ., 2011. Detection of a Cis [corrected] eQTL controlling BCMO1 gene expression leads to the identification of a QTG for chicken breast meat color. *PLoS One*.

TABLEAUX ET FIGURES

Tableau 1. Nombre de SNP parmi les SNP discriminant des lignées G/M et R+/R- identifiés dans les régions géniques classés selon leur impact fonctionnel sur la protéine

Effet du SNP sur l'ARN ou la protéine	G/M (510 577 SNP)		R+/R- (386 177 SNP)	
	nb SNP	nb Gènes	nb SNP	nb Gènes
non synonyme + épissage	1328 + 873	908 + 710	859 + 609	584 + 472
stop	28	28	18	17
SNP "codants"	2229	1646	1486	1070
3'UTR + 5'UTR	2006 + 241	1022 + 170	1303 + 155	673 + 118
promoteur 500 bases	2807	1587	2103	1190
SNP "régulateurs"	5054	2779	3561	1981
Nombre total	/	3327	/	2319
Nombre parmi les 15990 SNP communs aux deux populations	234 SNP et 171 gènes			

Tableau 2. Termes « Gene Ontology » enrichis associés aux SNP « codants » et « régulateurs » parmi les SNP discriminant des lignées G/M et R+/R-

	G/M	R+/R-
SNP "codants" Et "régulateurs"	Meiosis Carboxylic acid metabolic process* Purine ribonucleoside triphosphate metabolic process Indole derivative metabolic process Regulation of glycogen biosynthetic process* tRNA aminoacylation Aspartate catabolic process*	Meiosis ATP metabolic process* Macrophage activation Response to host immune response RNA splicing Regulation of synaptogenesis Glycogen metabolic process*
SNP "codants" Et "régulateurs" communs aux deux populations	Negative regulation of protein metabolic process Positive regulation of protein modification process Negative regulation of cell differentiation process Fatty acid biosynthetic process* Positive metabolic process* RNA splicing Monocarboxylic acid metabolic process*	

- : Termes associés aux métabolismes de l'énergie

Figure 1. Mutations dans le gène ACACA

