Y A-T-IL DE L'EMPREINTE CHEZ LA POULE ?

Frésard Laure[1], Leroux Sophie[1], Gourichon David[2], Dehais Patrice[1], Servin Bertrand [1], San Cristobal Magali[1], Marsaud Nathalie[3], Beaumont Catherine[4], Zerjal Tatiana[5], Kaiser Pete[6], Lagarrigue Sandrine[7], Vignal Alain[1], Morisson Mireille[1], Pitel Frédérique[1]

[1] INRA-ENVT - UMR444 Laboratoire de Génétique Cellulaire - 31326, Castanet-Tolosan ;
[2] INRA - PEAT Pôle d'Expérimentation Avicole de Tours - 37380, Nouzilly ;
[3] INRA - GeT-PlaGe Genotoul - 31326, Castanet-Tolosan ;
[4] INRA - UR83 Recherche Avicoles - 37380, Nouzilly ;
[5] INRA-AgroParisTech - UMR1313 Génétique Animale et Biologie Intégrative - 78350, Jouy ;
[6] The Roslin Institute and R(D)SVS, University of Edinburgh, United Kingdom;
[7] INRA-Agrocampus Ouest, UMR1348 PEGASE, Rennes, France

frederique.pitel@toulouse.inra.fr

RÉSUMÉ

La question de l'évolution de l'empreinte chez les vertébrés et de son existence chez les oiseaux est souvent évoquée dans la littérature mais aucune réponse définitive n'a été donnée pour le moment. L'empreinte génomique parentale est une modification épigénétique qui conduit à l'expression dépendante de l'origine parentale chez certains gènes. Les "effets réciproques" observés dans les espèces en croisement (différences de phénotype en fonction du sens de croisement entre lignées) résultent notamment de gènes liés au sexe, d'effets maternels ou de la transmission de l'ADN mitochondrial, mais une part de ces effets pourrait être expliquée par ce phénomène d'empreinte. L'empreinte génomique a été démontrée chez les mammifères euthériens et les marsupiaux mais jamais chez les monotrèmes ou les oiseaux. Jusqu'à présent, seule l'expression de quelques gènes, orthologues de gènes soumis à empreinte chez les mammifères, a été analysée chez la poule, et les résultats convergeaient vers la non-existence d'empreinte dans cette espèce. Des QTL soumis à empreinte ont pourtant été découverts chez la poule. Tandis que certains d'entre eux ont été invalidés, d'autres apparaissent comme des pistes sérieuses, avec l'utilisation d'approches méthodologiques et de dispositifs expérimentaux adaptés. L'objectif principal de notre projet est de détecter des gènes pour lesquels on observe des variations d'expression en fonction de l'allèle porté : l'expression allèle-spécifique d'une part, et l'expression dépendante de l'origine parentale d'autre part, démontrant l'existence potentielle du phénomène d'empreinte chez les oiseaux. L'ensemble du génome a été criblé à partir du transcriptome total d'embryons issus de séquençage haut-débit afin de répondre à cette question ; deux lignées de poules ont été utilisées, les plus consanguines et génétiquement distantes possibles, pour être capable d'identifier l'origine parentale des allèles observés. Deux familles issues de croisements réciproques ont été produites et les transcrits de 20 embryons de 4,5 jours ont été séquencés sur 6 lignes HiSeq 2000. Environ 200Gb de données ont été générés et leur analyse a permis la détection de 96 SNP candidats.

ABSTRACT

Does genomic imprinting exist in chickens?

Our project deals with the question of the imprinting evolution in vertebrates and its existence in birds, often evoked in the literature, but not definitely answered yet. Genomic imprinting is an epigenetic modification leading to parent-of-origin-specific expression of several genes. It has been observed in eutherians mammals and marsupials, but not in monotremes or birds. So far, the allelic expression of several genes orthologous to mammalian imprinted ones has been analyzed in chicken, without any reliable evidence of imprinting in this species. Several imprinted QTL were published in poultry; whereas some of them may finally be considered as not relevant to genomic imprinting, others appeared to be consistent, when using appropriate animal design and methodology. Our main objectives are to detect genes for which variation in expression is observed according to the allele, either because of an allele-specific expression or a parent-of-origin dependent expression. We thus screen the entire genome for allele-specific differential expression on whole embryonic transcriptomes by using high-throughput sequencing.

Two chicken lines were used, as inbred as possible and as genetically distant as possible, to unquestionably identify the parental origin of each observed haplotype. Two families were produced, coming from two reciprocal crosses. Transcripts from 20 embryos (4.5 days) have been tagged and sequenced through 6 HiSeq2000 lanes. About 200 Gb have been generated and their analysis allowed detection of 96 candidate SNPs.

INTRODUCTION

L'empreinte génomique parentale est l'expression de certains gènes de manière dépendante de l'origine parentale des allèles (Da Rocha *et al.*, 2004). Ce phénomène a été décrit chez les mammifères, les plantes et certains insectes, mais jamais chez les oiseaux. La principale théorie proposée pour expliquer ce phénomène est le conflit parental, qui soutient que les gènes contrôlant de près ou de loin l'allocation de ressources à l'embryon seraient sujets à une expression biaisée en fonction de l'origine parentale, le père privilégiant la croissance de ses descendants, la mère limitant l'utilisation de ses ressources pour préserver ses futures portées (Haig *et al.*, 1991, Moore *et al.*, 1991). Selon cette théorie, l'empreinte parentale devrait être restreinte aux organismes dont les ressources maternelles peuvent affecter directement l'expression des gènes de l'embryon. Il serait donc peu probable d'observer ce phénomène chez les animaux ovipares (Iwasa, 1998). Par ailleurs, des QTL soumis à empreinte ont été découverts chez la poule et la caille (De Koning *et al.*, 2002, Minvielle *et al.*, 2005, Tuiskula-Haavisto *et al.*, 2004, Tuiskula-Haavisto *et al.*, 2007), parmi lesquels des pistes sérieuses ont été mises en évidence grâce à des approches méthodologiques adaptées et des modèles permettant d'éviter des fausses détections (Rowe *et al.*, 2009). Jusqu'à présent, seule l'expression de quelques gènes connus pour être soumis à empreinte chez les mammifères a été analysée chez la poule (Koski *et al.*, 2000, Nolan *et al.*, 2001, O'neill *et al.*, 2000, Shin *et al.*, 2010, Wang *et al.*, 2005, Yokomine *et al.*, 2001, Yokomine *et al.*, 2005). Les études réalisées concluent pour la plupart d'entre elles à une expression biallélique des gènes testés et donc à la non-existence d'empreinte dans cette espèce. Aucune conclusion définitive n'a été donnée pour le moment. Par ailleurs, il est aujourd'hui possible d'étudier l'empreinte génomique à l'échelle du transcriptome par *RNA-sequencing* (Gregg *et al.*, 2010, Wang *et al.*, 2011). Ceci permet en particulier de s'interdire tout *a priori* dans le choix des zones à cibler. Cette méthodologie a été choisie pour tenter de répondre à la question brûlante de l'existence de l'empreinte chez la poule.

1. MATERIELS ET METHODES

1.1. Procédures expérimentales

Des embryons F1 ont été générés par croisement réciproque de poules (6 mâles et 6 femelles par sens de croisement, Figure 1) des lignées 6 (Bumstead *et al.*, 1988) et R^- (Bordas *et al.*, 1992) éloignées génétiquement et assez consanguines : le choix a été réalisé à partir du génotypage d'individus de différentes lignées sur une puce 57K, qui a permis d'estimer le taux de consanguinité des lignées candidates (0.97 pour la lignée 6 et 0.87 pour la lignée R^-) et leurs distances génétiques respectives (distances de Reynolds (Reynolds *et al.*, 1983), Figure 2).
Une extraction ADN-ARN a été réalisée sur les embryons à 4,5 jours (stade 24-25) avec le kit AllPrep DNA/RNA Mini Kit (Qiagen), selon les recommandations du fournisseur.
L'ADN des parents a été obtenu à partir d'échantillons de sang par une méthode d'extraction rapide (Roussot *et al.*, 2003).
Les banques de séquençage ont été préparées par la technique Illumina adaptée au RNAseq (seuls les ARNm, polyA, sont sélectionnés). Les échantillons ont été "tagués" (identifiés par l'addition en extrémité d'un court polymère d'ADN de séquence connue) pour permettre leur identification ultérieure, puis séquencés en triplicata sur un séquenceur HiSeq 2000 (Illumina), en randomisant au maximum leur position dans les différentes lignes de séquence. En tout, 6 lignes de séquençage ont été utilisées.
De plus, l'ADN des parents des F1 a été séquencé sur 4 lignes afin de permettre la détection des SNP discriminants les deux lignées.

1.2. Analyse des données

Sauf indication contraire, toutes les analyses ont été réalisées grâce à des scripts perl, python et R écrits sur mesure.

Identification de SNP discriminants les deux lignées
Les lectures issues du séquençage d'ADN des parents ont été alignées sur la dernière version du génome de référence de la poule (Galgal4) par le programme bwa (Li *et al.*, 2009). A partir de ces données, un comptage allélique tout génome a été réalisé. Ces données de comptage ont permis de sélectionner uniquement les locus discriminant les parents, c'est-à-dire pour lesquels les deux parents étaient homozygotes pour deux allèles différents.

Identification de SNP à expression biaisée
Le logiciel Tophat (Trapnell *et al.*, 2009) a été utilisé pour aligner les séquences issues du cDNA des embryons F1 sur la dernière version du génome de la poule (Galgal4), en lui imposant de ne garder que les lectures n'ayant qu'un seul alignement possible sur le génome. Ces données d'alignement ont ensuite été regroupées en fonction de leur sens de croisement d'origine pour la suite des analyses. Des comptages alléliques ont alors été réalisés, puis les données ont été filtrées selon plusieurs critères : seules les positions biallélliques, avec un nombre de lectures supérieur à 10 dans les deux sens de croisement et présentant un ratio d'expression inversé entre les croisements ont été retenues. Pour tester le biais d'expression allélique, un test exact de Fisher a été effectué sur tous les locus exprimés. Le seuil de significativité a été fixé à un FDR<0.05 (False Discovery Rate) après utilisation du package R "qvalue" (Dabney *et al.*, 2009).
Enfin, les données filtrées ont été croisées avec les positions discriminant les lignées préalablement détectées.

2. RESULTATS ET DISCUSSION

Au total, 2 298 622 SNP discriminants (données brutes) ont été détectés entre les parents des deux lignées à partir de leurs séquences d'ADN génomique, ce qui correspond en moyenne à 2,1 ±0,7 SNP/kb sur tout le génome. Une étude de détection de SNP par re-

séquençage d'ADN de lignées de poules sélectionnées de manière divergente sur le poids juvénile a permis d'identifier 1 338 437 SNPs dont l'allèle diffère totalement de celui de la référence (Red Jungle Fowl) (Marklund *et al.*, 2010). Ceci souligne que les lignées utilisées dans notre étude sont assez éloignées génétiquement et augmenteront les possibilités de discriminer les allèles observés en fonction de leur origine parentale.

Les séquences obtenues à partir du transcriptome de tous les embryons de 4,5 jours couvrent 56% du génome du poulet. Un critère de profondeur minimale de 10 lectures a été décidé pour réaliser le calcul sur les données de tous les embryons réunis. Ceci est un peu inférieur à la couverture de 58,5% obtenue en transcriptome humain à partir de 15 lignées cellulaires de différents tissus (Djebali *et al.*, 2012) à profondeur égale.

Parmi les 200 Gb de séquences issues du RNAseq, 298 776 025 pb sont exprimées dans les deux sens de croisement, au sens où au moins une lecture a été observée dans chaque sens de croisement. Parmi ces positions, 40% (117 219 312) ont une profondeur supérieure à 10 dans les deux sens de croisement et parmi ces dernières, 4% (5 039 615 positions) sont bialléliques. Un test exact de Fisher a été réalisé sur ces positions. En limitant le FDR à 5%, 2% des positions restantes apparaissent comme significatives, c'est-à-dire qu'à ces locus, les profils d'expression entre les deux croisements sont différents. Seules les positions présentant un ratio allélique inversé entre les deux sens de croisement (Fold Change >2.5, l'allèle majoritaire dans un sens de croisement devient minoritaire dans l'autre) étaient potentiellement candidates. On n'attend pas forcément d'extinction complète d'un des allèles puisque l'étude est conduite sur des embryons entiers, donc un mélange de tissus. Ce critère a permis la sélection de 4 849 locus. Enfin, uniquement les positions où l'origine parentale des allèles peut être connue sont à retenir ici. Les résultats ont donc été croisés avec les SNP discriminants les parents des deux lignées utilisées (préalablement détectés). En fin d'analyse, 96 locus sont candidats et correspondent au profil d'empreinte recherché. Ces SNP sont situés en majorité dans des introns de gènes (79% d'entre eux) (Figure 3), les autres étant répartis dans des exons et des UTR (Untranslated Region, en amont ou en aval de la région codante des gènes). Ce positionnement intronique n'est pas incompatible avec un phénomène d'empreinte génétique. En effet, il a été montré, chez la souris en particulier, que des rétrogènes soumis à empreinte pouvaient se situer dans des introns de gènes porteurs ne présentant pas d'empreinte (Monk *et al.*, 2011). Il a de plus été démontré que chez certains gènes qui ont des transcrits alternatifs soumis à empreinte, les sites de polyadénylation alternatifs (en 3'-UTR des gènes) peuvent être contrôlés de manière dépendante de l'origine parentale (Cowley *et al.*, 2012, Wood *et al.*, 2008).

Les locus mis en évidence lors de cette étude ne semblent pas se rassembler en cluster comme c'est le cas chez les mammifères (Edwards *et al.*, 2007). On peut d'une part émettre l'hypothèse que la profondeur des séquences n'est pas suffisante pour affiner l'analyse du système d'organisation de ces locus candidats. D'autre part, il est aussi possible que ces clusters de gènes soumis à empreinte soient inexistants chez la poule et que d'autres mécanismes de régulation aient été mis en place au cours de l'évolution.

Des biais d'analyse inhérents au séquençage haut-débit peuvent générer de faux positifs (Deveale *et al.*, 2012, Kelsey *et al.*, 2012) : des allèles peuvent être légèrement sous-représentés lors de la préparation des banques, l'alignement des lectures peut être légèrement biaisé en fonction de l'allèle de la séquence de référence, des SNP observés peuvent se révéler artéfactuels… Les locus que nous avons mis en évidence sont donc en cours de validation, par séquençage Sanger dans une première approche (Figure 4) et pyroséquençage, technologie plus précise pour la quantification.

CONCLUSION

Le séquençage ARN en haut débit s'avère être un outil efficace pour repérer des phénomènes non encore étudiés à l'échelle du génome. En s'affranchissant de l'étude gène par gène, il est ainsi possible de détecter des phénomènes se rapportant à de l'empreinte chez des animaux non mammifères. Des validations sont nécessaires pour pallier les possibles biais d'analyse (induits dès la construction des librairies jusqu'à l'analyse des séquences pouvant engendrer des erreurs systématiques de quantification dans l'analyse statistique) soulevés par cette méthode et pour conclure de manière plus précise sur l'existence potentielle d'empreinte génomique chez la poule. Ces résultats pourraient ainsi signifier que ce type d'expression aurait été mis en place aussi chez des espèces vertébrées non mammifères. Des régulations et modalités de mise en place différentes de celles existant chez les mammifères ne seraient en aucun cas à exclure, ce qui permettrait peut-être de ne plus évoquer l'empreinte mais les empreintes.

REMERCIEMENTS

Ce travail a été réalisé grâce au soutien de l'ANR (programme EpiBird ANR-009-GENM-004). LF bénéficie d'une bourse de thèse Région Midi-Pyrénées / DGA INRA.

REFERENCES BIBLIOGRAPHIQUES

Bordas A, Tixier-Boichard M et Merat P, 1992. Br Poult Sci, (33), 741-54.
Bumstead N et Barrow PA, 1988. Br Poult Sci, (29), 521-9.

Cowley M, Wood AJ, Bohm S, Schulz R et Oakey RJ, 2012. Nucleic Acids Res, (40), 8917-26.
da Rocha ST et Ferguson-Smith AC, 2004. Curr Biol, (14), R646-9.
Dabney A et Storey JD, 2009. R package, http://CRAN.R-project.org/package=qvalue
de Koning D-J, Bovenhuis H et van Arendonk JAM, 2002. Genetics, (161), 931-938.
DeVeale B, van der Kooy D et Babak T, 2012. PLoS Genet, (8), e1002600 EP -.
Djebali S *et al.*, 2012. Nature, (489), 101-108.
Edwards CA et Ferguson-Smith AC, 2007. Current Opinion in Cell Biology
Nucleus and Gene Expression, (19), 281-289.
Gregg C, Zhang J, Weissbourd B, Luo S, Schroth GP, Haig D et Dulac C, 2010. Science, (329), 643-648.
Haig D et Graham C, 1991. Cell, (64), 1045-6.
Iwasa Y, 1998. In: Current Topics in Developmental Biology, R. A. P. a. G. P. Schatten, Academic Press
Kelsey G et Bartolomei MS, 2012. PLoS Genet, (8), e1002601.
Koski LB, Sasaki E, Roberts RD, Gibson J et Etches RJ, 2000. Molecular Reproduction and Development, (56), 345-352.
Li H et Durbin R, 2009. Bioinformatics, (25), 1754-60.
Marklund S et Carlborg O, 2010. BMC Genomics, (11), 665.
Minvielle F, Kayang BB, Inoue-Murayama M, Miwa M, Vignal A, Gourichon D, Neau A, Monvoisin JL et Ito S, 2005. BMC Genomics, (6), 87.
Monk D, Arnaud P, Frost JM, Wood AJ, Cowley M, Martin-Trujillo A, Guillaumet-Adkins A, Iglesias Platas I, Camprubi C, Bourc'his D, Feil R, Moore GE et Oakey RJ, 2011. Nucleic Acids Res, (39), 4577-86.
Moore T et Haig D, 1991. Trends in Genetics, (7), 45-49.
Nolan CM, Killian JK, Petitte JN et Jirtle RL, 2001. Dev Genes Evol., (211), 179-183.
O'Neill MJ, Ingram RS, Vrana PB et Tilghman SM, 2000. Dev Genes Evol, (210), 18-20.
Reynolds J, Weir BS et Cockerham CC, 1983. Genetics, (105), 767-79.
Roussot O, Feve K, Plisson-Petit F, Pitel F, Faure JM, Beaumont C et Vignal A, 2003. Genet Sel Evol, (35), 559-72.
Rowe SJ, Pong-Wong R, Haley CS, Knott SA et de Koning DJ, 2009. Genet Sel Evol, (41),
Shin S, Han JY et Lee K, 2010. Poultry Science, (89), 948-955.
Trapnell C, Pachter L et Salzberg SL, 2009. Bioinformatics, (25), 1105-11.
Tuiskula-Haavisto M, De Koning D, Honkatukia M, Schulman NF, Mâki-Tanila A et Vilkki J, 2004. Genet Res, (84), 57-66.
Tuiskula-Haavisto M et Vilkki J, 2007. Cytogenet Genome Res, (117), 305-12.
Wang G, Yan B, Deng X, Li C, Hu X et Li N, 2005. Sci China C Life Sci., (48), 187-194.
Wang X, Soloway PD et Clark AG, 2011. Genetics, (189), 109-122.
Wood AJ, Schulz R, Woodfine K, Koltowska K, Beechey CV, Peters J, Bourc'his D et Oakey RJ, 2008. Genes Dev, (22), 1141-6.
Yokomine T, Kuroiwa A, Tanaka K, Tsudzuki M, Matsuda Y et Sasaki H, 2001. Cytogenetic and Genome Research, (93), 109-113.
Yokomine T, Shirohzu H, Purbowasito W, Toyoda A, Iwama H, Ikeo K, Hori T, Mizuno S, Tsudzuki M, Matsuda Y, Hattori M, Sakaki Y et Sasaki H, 2005. Genome Res, (15), 154-65.

Figure 1. Dispositif expérimental.
Un croisement réciproque est réalisé entre deux lignées distantes génétiquement et le plus consanguines possible. En cas d'empreinte génomique, les embryons expriment uniquement l'allèle d'un de leurs parents (cas d'empreinte à expression paternelle ici). La détection est possible grâce aux polymorphismes différenciant les deux lignées utilisées

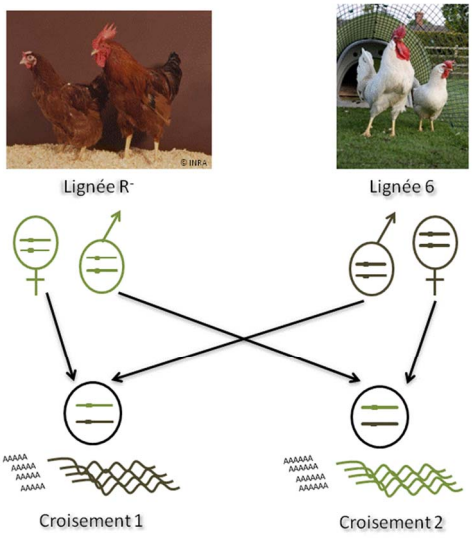

Figure 2. Heatmap de la matrice de similarité des lignées testées sur la puce 57K.
Le dendrogramme ne diffère pas de celui obtenu à partir des distances génétiques de Reynolds. Les couleurs correspondent au degré de consanguinité des individus testés (rouge correspondant à consanguin). Les lignées 6 et R⁻ ont été choisies pour être les plus consanguines et éloignées génétiquement possible. Elles sont indiquées par les flèches.

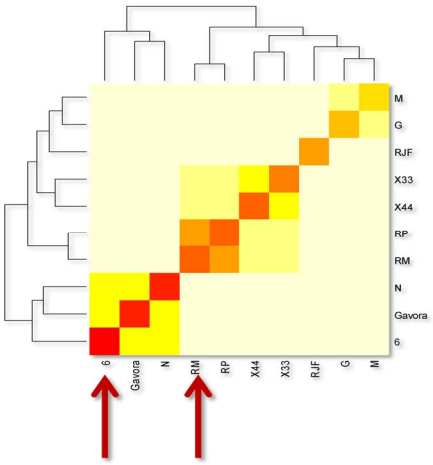

Figure 3. Observation des SNP candidats par Integrative Genomics Viewer (IGV).
La première ligne de la fenêtre de visualisation correspond à la profondeur des lectures sur le génome de référence. Les zones « les plus couvertes » correspondent donc aux exons des gènes. Des lectures sont observées entre ces exons mais correspondent pourtant à la localisation des introns, partiellement exprimés. Les régions candidates sont majoritairement introniques.

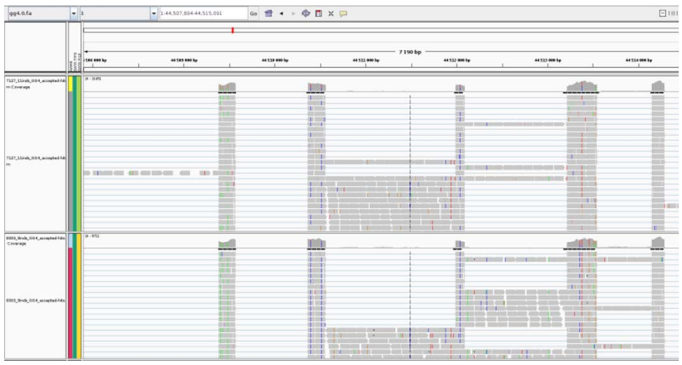

Figure 4. Confirmation de certains SNP candidats par séquençage Sanger.
A droite et à gauche se trouvent respectivement les données du croisement 1 et du croisement 2 sur un individu (vérifié sur respectivement 12 et 8 individus dans les croisements 1 et 2). La ligne du haut correspond aux données génomiques (deux pics superposés, SNP biallélique) tandis que la ligne du bas représente le cDNA : extinction d'un allèle dans le sens de croisement 1 (pic éteint) et surexpression de cet allèle dans le sens de croisement 2 (pic au dessus).