ANALYSE DU METHYLOME CHEZ L'EMBRYON DE POULET PAR SEQUENÇAGE HAUT-DEBIT

Mersch Marjorie[1], Noirot Céline[2], Leroux Sophie[1], Esquerré Diane[3], Gourichon David[4], Lefort Gaëlle[2], Salin Gérald[3], Djebali Sarah[1], Frésard Laure[5], Foissac Sylvain[1], Pitel Frédérique[1]

[1]GenPhySE, Université de Toulouse, INRA, INPT, ENVT, 31326 CASTANET-TOLOSAN ;
[2]INRA, UR875 MIAT, 31326 CASTANET-TOLOSAN ;
[3]INRA, GeT-PlaGe Genotoul, 31326 CASTANET-TOLOSAN ;
[4]INRA, PEAT, 37380 NOUZILLY ;
[5]Department of Pathology, Stanford University, Stanford, CA 94305, USA

marjorie.mersch@inra.fr

RÉSUMÉ

L'étude des causes moléculaires des variations phénotypiques observées chez les espèces d'élevage passe généralement par l'analyse de la variabilité de la séquence d'ADN. Récemment, de nombreuses études décrivent le rôle important de l'environnement dans la régulation de l'expression des gènes, notamment via des phénomènes épigénétiques. Afin de caractériser une des marques épigénétiques chez le poulet, la méthylation de l'ADN, un dispositif de 8 embryons (4 mâles et 4 femelles) issus du croisement réciproque de deux lignées a été utilisé. À partir d'une technique d'analyse par séquençage à haut-débit des marques de méthylation, un pipeline d'analyse bioinformatique a été développé. Un taux de méthylation plus faible dans les régions promotrices que dans les régions géniques et intergéniques a été observé. Grâce aux données transcriptomiques également disponibles sur ces embryons, nous avons confirmé que le niveau d'expression des gènes est plus fort quand les régions promotrices sont peu méthylées, et inversement. La possible interaction du fond génétique ou du sexe avec le niveau de méthylation, et son lien avec les niveaux d'expression chez les 8 embryons est en cours d'analyse. Ces analyses vont être étendues dans le cadre du projet ANR ChickStress qui étudie la réponse du méthylome à des variations de température et à l'utilisation d'un aliment de faible valeur énergétique chez la poule.

ABSTRACT

Methylome analysis through high-throughput sequencing in chicken embryos

Investigating the molecular causes of phenotypic variations observed in farmed species usually involves the study of the variability in the DNA sequence. Recently, many studies have described the important role of the environment in the regulation of gene expression, particularly through epigenetic mechanisms. In order to characterize one of these epigenetic marks in chicken, DNA methylation, we have used 8 embryos (4 males and 4 females) from the reciprocal cross of two lines. We developed a bioinformatic analysis pipeline to analyze methylation marks from high-throughput sequencing. We have observed a lower methylation rate in the promoter regions than within the genes and intergenic regions. Thanks to the transcriptomic data also available on these embryos, we have shown that the gene expression level was higher when the promoter regions were poorly methylated, and vice versa. The possible interaction of the genetic background or sex with the methylation level and the link with expression level in these 8 embryos is being analyzed. These analyses will be extended as part of the ANR ChickStress project, which studies the methylome response to changes in temperature and the use of a low-energy diet in chicken.

INTRODUCTION

Un des objectifs majeurs de la recherche en agronomie est d'étudier les causes moléculaires des variations phénotypiques observées dans les espèces d'élevage. Celles-ci sont dues à des facteurs génétiques et/ou environnementaux. L'étude des phénomènes épigénétiques potentiellement impliqués constitue aujourd'hui un moyen prometteur pour mieux comprendre ces effets. Il est maintenant clair que l'environnement module l'expression des gènes via ces phénomènes épigénétiques, participant ainsi à la plasticité phénotypique (Mazzio *et al.*, 2012).

La variation des phénotypes est induite, en partie, par des fluctuations dans la régulation de l'expression des gènes. Ceci permet de moduler l'activité des cellules d'un organisme en fonction de leur environnement. Parmi les acteurs de cette régulation, il existe des mécanismes qui n'affectent pas la séquence d'ADN mais qui peuvent être transmis par la mitose ou la méiose : ce sont les phénomènes épigénétiques (Feil et Fraga, 2011). Ces mécanismes sont matérialisés par des marques biochimiques, en partie réversibles, sur l'ADN ou sur les protéines qui le structurent : les histones. La méthylation de l'ADN constitue un des principaux mécanismes décrits à ce jour (Egger *et al.*, 2004). Il s'agit d'une modification chimique des cytosines (C) en 5-méthylcytosine par l'ajout d'un groupement méthyle (CH_3) qui affecte ainsi la régulation de l'expression des gènes, en modifiant notamment l'accessibilité de la chromatine de certains promoteurs (Deaton et Bird, 2011). Chez les Vertébrés, les 5-méthylcytosines sont majoritairement observées dans les dinucléotides CG. Ce processus est catalysé par une famille d'enzymes appelées les DNA méthyltransferases.

Le méthylome est défini comme l'ensemble des modifications de l'état de méthylation le long du génome (Feinberg, 2001). Sa caractérisation permet de faire le lien entre ces marques biochimiques et le niveau d'expression des gènes à l'échelle du génome et donc de mieux comprendre les mécanismes de régulation de l'expression des gènes. Son étude consiste en l'observation du taux de méthylation de chaque cytosine dans les dinucléotides CG. Ce taux de méthylation correspond à la proportion de cytosines méthylées à une position donnée, observée en général à partir d'un ensemble de molécules d'ADN provenant de plusieurs cellules.

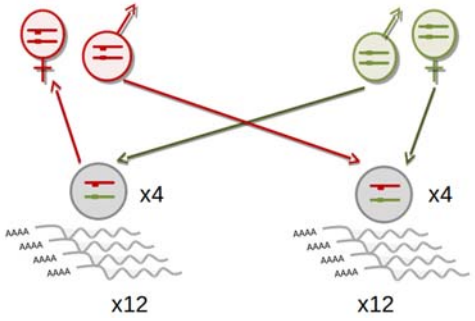
L'étude du méthylome est facilitée aujourd'hui par les nouvelles technologies de séquençage. L'objectif de cette étude est de caractériser le méthylome de l'embryon de poulet - stade pour lequel nous disposions également de données d'expression - et d'analyser les liens existant entre le taux de méthylation de l'ADN et le niveau d'expression des gènes le long du génome.

1. MATÉRIELS ET MÉTHODES

1.1 Dispositif d'étude

Les échantillons analysés sont issus d'un dispositif animal mis en place pour étudier l'empreinte génomique chez le poulet (Frésard *et al.*, 2014). Il s'agit d'un croisement réciproque de deux lignées de poule (4 parents) ayant produit 12 embryons (à 4,5 jours) avec 6 mâles et 6 femelles par sens de croisement (figure 1).

Figure 1. Croisement réciproque des deux lignées donnant 12 embryons par sens de croisement



Pour 8 embryons (4 mâles et 4 femelles) sur les 24, deux types de données sont produites, issues d'expériences de séquençage à haut-débit. Premièrement, des données produites par RNA-seq (séquençage du transcriptome) permettent d'attribuer à chaque gène un niveau d'expression dans chacun des embryons. Deuxièmement, des données produites par DNA-seq (séquençage du génome) identifient les variations au sein de la séquence d'ADN propre à chaque individu. Au cours de cette étude, une analyse complémentaire a été ajoutée : les motifs de méthylation ont été détectés par une technologie appelée *Whole Genome Bisulfite Sequencing* (WGBS).

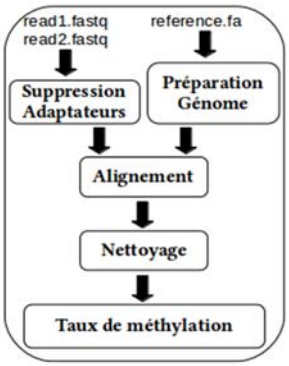
1.2 Séquençage WGBS

Nous avons utilisé les échantillons d'ADN produits pour le séquençage génomique classique. Le WGBS est une technique de séquençage qui repose sur un traitement chimique préalable de l'ADN génomique au bisulfite de sodium, permettant de distinguer les cytosines méthylées des autres (Krueger *et al.*, 2012). En effet, une cytosine méthylée n'est pas affectée par le traitement et reste une cytosine dans les lectures issues du séquençage, alors qu'une cytosine non méthylée est désaminée en uracile par le bisulfite de sodium, puis lue en tant que thymine lors du séquençage, après amplification. Cette conversion est détectée par séquençage haut-débit et apporte l'information de l'état de méthylation originel. Les données de séquence ont ensuite été analysées grâce à un pipeline (chaîne de traitement bioinformatique d'analyse des données de séquence).

1.3 Construction d'un pipeline

Le pipeline développé à l'aide de scripts Shell et R permet d'effectuer les analyses depuis le traitement initial des données de séquençage jusqu'aux analyses de méthylation (figure 2).

Figure 2. Schématisation des étapes du pipeline



La première étape du pipeline consiste à éliminer les parties non informatives des lectures, c'est-à-dire les adaptateurs moléculaires nécessaires au protocole expérimental lors du séquençage, avec le logiciel Trim Galore! (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore). Les séquences trop courtes à l'issue de cette étape sont éliminées. Un contrôle qualité est réalisé par l'outil FastQC (http://www.bioinformatics.babraham.ac.uk/projects/trim_fastqc). Cela permet de vérifier que les séquences obtenues ont toutes les qualités requises. Sur des données génomiques non traitées au bisulfite chez la poule, on s'attend à un pourcentage de C et de G d'environ 20 % et de T et de A d'environ 30 % (Hillier *et al.*, 2004).
Les nombreuses lectures obtenues peuvent provenir de n'importe quelle partie du génome. Il faut donc aligner chaque lecture sur le génome de référence de l'espèce pour obtenir sa position. Le logiciel Bismark (Krueger *et al.*, 2011) est classiquement utilisé dans ce type d'analyse. Le génome de référence utilisé pour cette étude est la version 5 de *Gallus gallus* du navigateur de génome *Ensembl*. Les séquences alignées sont ensuite filtrées pour éliminer les duplicats PCR (dus à l'étape d'amplification PCR des fragments ADN avant séquençage) en utilisant l'option *rmdup* de SAMtools (Li *et al.*, 2009).
L'étape suivante consiste à calculer l'état de méthylation moyen de chaque position du génome portant une cytosine. Pour cela, le *package* R MethylKit (Akalin *et al.*, 2012) extrait toutes les positions de C dans les dinucléotides CG à partir des lectures. Le taux de méthylation d'un dinucléotide CG est la proportion de C observés dans les lectures alignées à la position correspondante, donnée par la formule : C/(C+T).
Néanmoins, la variabilité de séquence due au polymorphisme individuel, c'est-à-dire à la présence de *Single Nucleotide Polymorphism* (SNP), peut fausser le calcul de méthylation. À partir des données génomiques non traitées (DNA-seq), l'outil GATK (*Genome Analysis Toolkit* ; DePristo *et al.*, 2011) identifie les positions pour lesquelles le taux de méthylation n'est pas fiable. Ces positions sont ensuite éliminées des positions détectées précédemment.

Les taux de méthylation à chaque position sont obtenus sur les 2 brins (sens et antisens). De nombreuses études ont montré que la méthylation est symétrique chez les Vertébrés (Reik *et al.*, 2011), ainsi nous additionnons les informations de méthylation (couverture de la position et nombre de C et de T en cette position) entre les deux brins afin d'obtenir un taux de méthylation en une position CG.

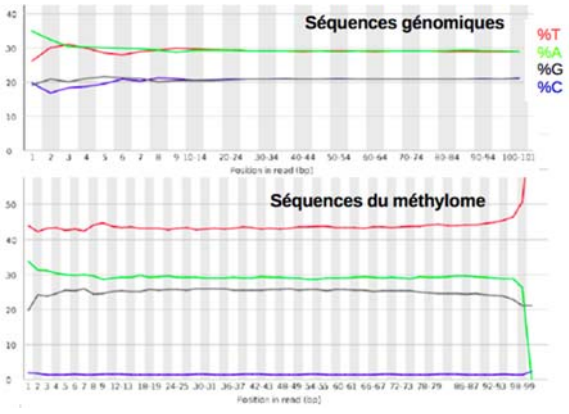
Afin de comparer le niveau de méthylation entre les embryons, une étape préliminaire de normalisation est nécessaire. En effet, il s'agit de compenser les variations expérimentales du nombre total de lectures produites par embryon (profondeur de séquençage). Cette étape est réalisée dans le *package* MethylKit. Ensuite, seules les positions communes à tous les embryons sont retenues.

2. RÉSULTATS ET DISCUSSION

2.1 Caractéristiques des données obtenues

Les résultats présentés sont une moyenne entre les 8 embryons. Le nombre de lectures varie de 134 millions à 198 millions par échantillon, pour une moyenne de 161 millions de lectures par embryon. Ces variations, dues en particulier à la quantité d'ADN de départ et à l'efficacité de l'amplification, sont corrigées par la normalisation. Après filtre Trim Galore!, il reste 156 millions de lectures par embryon. Lorsque le traitement au bisulfite a bien fonctionné, la proportion de C doit diminuer au profit de celle de T. Sur la figure 3, T augmente à plus de 40 % et C diminue à 1 %. Cela signifie que le traitement au bisulfite a correctement remplacé les C non méthylés par des T et que la majorité des C présents dans le génome sont non méthylés.

Figure 3. Visualisation de la composition en bases des lectures d'un embryon par FastQC en fonction de la position le long du génome, sur des séquences génomiques et sur des séquences du méthylome.



À l'issue de l'étape d'alignement, 71,6 % des lectures sont alignées en une position unique, soit environ 112 millions de lectures. Puis vient la suppression des duplicats de PCR au cours de laquelle 15 % de lectures sont éliminées (95 millions de lectures au final).

Un contrôle interne est réalisé pour confirmer la fiabilité du traitement au bisulfite. Hors dinucléotides CG, les cytosines sont non méthylées et sont donc transformées en T après traitement au bisulfite. Sur les embryons, 99,33 % des C hors CG sont converties.

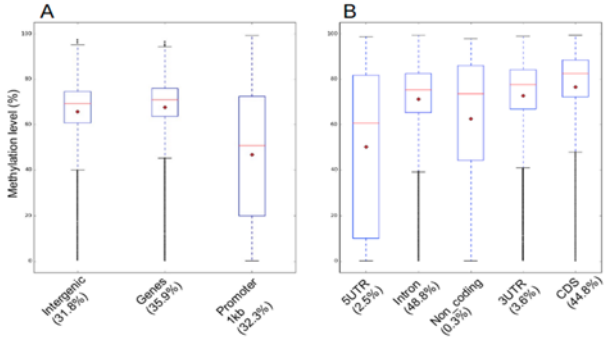
Après toutes ces étapes de traitement, le nombre de positions détectées est de 20,5 millions (sens et antisens) sur les 22 millions que compte le génome de référence. Une fois l'information des brins sens et antisens ramenée aux dinucléotides, 12,5 millions de positions CG avec un taux de méthylation associé sont détectées. La moyenne de méthylation sur tout le génome est d'environ 66,5 %.

Une fois la normalisation réalisée, il s'agit d'obtenir exactement les mêmes positions génomiques pour tous les embryons, chaque embryon ayant un taux de méthylation spécifique. Le nombre de SNP détectés par GATK sur l'ensemble du génome est de 9,7 millions. Après élimination des SNP correspondant à des positions retenues pour l'analyse de méthylation, 5,7 millions de CG sont analysables sur l'ensemble des embryons.

2.1 Caractéristiques fonctionnelles du méthylome

Les données obtenues permettent d'analyser la variation des taux de méthylation en fonction des régions du génome. Celui-ci est découpé en plusieurs régions fonctionnelles : promotrices, régulatrices, codantes ou non, intergéniques (annotation *Ensembl*, http://www.ensembl.org/Gallus_gallus/Info/Index). Il faut rechercher les positions de C au sein de ces régions afin d'associer un taux de méthylation à une région génomique particulière. La distribution du taux de méthylation en fonction de l'annotation est représentée sur la figure 4.

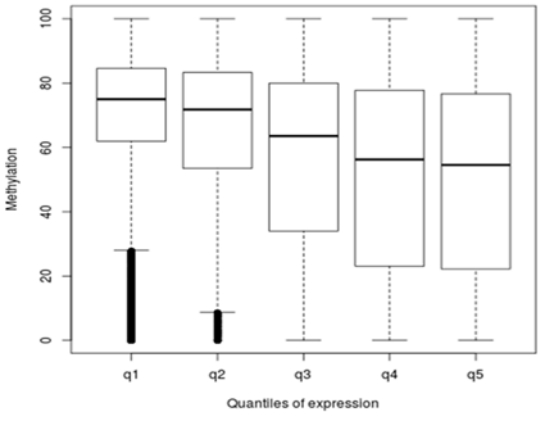
Figure 4. Distribution du taux de méthylation de chaque position en fonction du type de régions observées le long du génome. La région promotrice est arbitrairement définie comme étant 1kb en amont du site de début de transcription.



Sur la figure 4A, les taux de méthylation détectés sont répartis dans trois régions distinctes : 32,3 % des positions sont dans les régions promotrices, qui ont un taux de méthylation plus faible que les deux autres régions. Les différences de méthylation observées en fonction de la nature de la région génomique sont statistiquement significatives ($p = 2.10^{-16}$, test de Kruskal-Wallis). De plus, au sein des régions géniques (figure 4B) la répartition des taux est variable (analyses statistiques en cours). Dans les régions codantes, le taux de méthylation est le plus élevé. Il faut noter que la répartition dans les régions promotrices est très hétérogène (figure 4A : boxplot des régions promotrices de 20 % à 70 % de méthylation) au contraire des deux autres régions (figure 4A : boxplot de 60 % à 75 % de méthylation). Cette distribution particulière est liée à la quantité importante de promoteurs avec de faibles taux de méthylation (0-10%) ou de forts taux de méthylation (90-100%). Ce taux de méthylation plus faible dans les promoteurs est retrouvé dans le foie, le muscle et le poumon de poule (Li *et al*, 2011; Li *et al.*, 2015). Plusieurs études ont montré que les régions peu méthylées sont généralement dans les promoteurs chez les mammifères (voir Suzuki et Bird, 2008). Le promoteur, en amont du gène, est la région à partir de laquelle s'effectue l'initiation de la transcription de l'ADN en ARN.

Pour relier ce taux de méthylation au niveau d'expression des gènes, nous avons utilisé les données issues de RNA-seq (transcriptome) des mêmes échantillons.

Figure 5. Distribution du score de méthylation dans les promoteurs en fonction de quantiles d'expression des gènes.



Sur la figure 5, le niveau d'expression est divisé en 5 quantiles, le 1er quantile représentant le niveau d'expression le plus faible tandis que le 5e quantile représente le niveau d'expression le plus fort. Lorsque le promoteur est fortement méthylé, le niveau d'expression des gènes est globalement faible et inversement ($p = 2.10^{-16}$, test de Kruskal-Wallis), ce qui confirme notamment des résultats obtenus chez le

poulet : Li *et al.* ont démontré une corrélation négative entre la méthylation de l'ADN dans les régions promotrices et l'expression des ARN messagers sur des tissus pulmonaires prélevés chez le poulet (Li *et al.*, 2015). Cela suggère un rôle de régulation transcriptionnelle de la méthylation de l'ADN (Jaenisch et Bird, 2003), en particulier via les propriétés d'ouverture de la chromatine liées aux ilôts CpG non méthylés (Deaton et Bird, 2011).

Un des objectifs de ce travail est également d'évaluer s'il existe des différences de méthylation selon le sexe de l'embryon ou encore selon le sens de croisement. Le réglage des critères les plus pertinents pour l'analyse statistique (ampleur de la différence, p-valeur associée à un test statistique...) est actuellement en cours.

CONCLUSION

La conception, le développement et l'application d'un pipeline bioinformatique nous a permis de caractériser la méthylation de l'ADN génomique chez l'embryon de poulet. Les données générées par séquençage haut-débit rendent un tel pipeline très utile pour l'analyse biologique. En effet, ce pipeline d'analyses sera adapté dans le cadre du projet ANR ChickStress qui étudie la réponse du méthylome à des variations de température et à l'utilisation d'un aliment de faible valeur énergétique chez la poule. Dans ce projet, la technologie utilisée analyse une sous-représentation du génome et non plus le génome entier. Ainsi, cet outil pourra analyser des données issues de deux types de séquençage, pour les programmes en cours chez la poule mais également pour d'autres espèces. D'autres analyses complémentaires sont en cours, notamment sur la question du lien entre la méthylation différentielle et l'expression différentielle.

RÉFÉRENCES BIBLIOGRAPHIQUES

Akalin, A.; Kormaksson, M.; Li, S.; Garrett-Bakelman, F.; Figueroa, M.; Melnick, A.; Mason, C. *Genome Biol.* **2012**, *13*, R87.

Deaton, A. M.; Bird, A. P. *Genes & Development* **2001**, *25*,1010–1022.

DePristo, M. A.; Banks, E.; Poplin, R.; Garimella, K. V; Maguire, J. R.; Hartl, C.; Philippakis, A. A.; del Angel, G.; Rivas, M. A.; Hanna, M.; McKenna, A.; Fennell, T. J.; Kernytsky, A. M.; Sivachenko, A. Y.; Cibulskis, K.; Gabriel, S. B.; Altshuler, D.; Daly, M. J. **2011**, *43*, 491–8.

Egger, G.; Liang, G.; Aparicio, A.; Jones, P. A. *Nature* **2004**, *429*, 457–463.

Feil, R.; Fraga, M. F. *Nat. Rev. Genet.* **2012**, *13*, 97–109.

Feinberg, A. P. *Nat. Genet.* **2001**, *27*, 9–10.

Frésard, L.; Leroux, S.; Servin, B.; Gourichon, D.; Dehais, P.; San Cristobal, M.; Marsaud, N.; Vignoles, F.; Bed'hom, B.; Coville, J. L.; Hormozdiari, F.; Beaumont, C.; Zerjal, T.; Vignal, A.; Morisson, M.; Lagarrigue, S.; Pitel, F. *Nucleic Acids Res.* **2014**, *42*, 3768–3782.

Hillier, LDW; International Chicken Genome Sequencing Consortium. *Nature* **2004**, *432*, 695–716.

Jaenisch, R.; Bird, A. *Nat. Genet.* **2003**, *33 Suppl*, 245–254.

Krueger, F.; Andrews, S. R. *Bioinformatics* **2011**, *27*, 1571–1572.

Krueger, F.; Kreck, B.; Franke, A.; Andrews, S. R. *Nat Methods* **2012**, *9*, 145–151.

Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. *Bioinformatics* **2009**, *25*, 2078–2079.

Li, J.; Li, R.; Wang, Y.; Hu, X.; Zhao, Y.; Li, L.; Feng, C.; Gu, X.; Liang, F.; Lamont, S. J.; Hu, S.; Zhou, H.; Li, N. *BMC Genomics* **2015**, *16*, 1–13.

Mazzio, E. A.; Soliman, K. F. A. *Epigenetics* **2012**, *7*, 119–130.

Reik, W.; Dean, W.; Walter, J. *Epigenetics* **2001**, *293*, 1089–1094.

Suzuki, M. M.; Bird, A. *Nature Reviews Genetics* **2008**, *9*, 465–476.

Tate, P. H.; Bird, A. P. *Curr. Opin. Genet. Dev.* **1993**, *3*, 226–231.