

```
52 chrom_merged <- colsplit(merged_df$rs, ':', c('chr', 'pos'))  
53 start_merged <- colsplit(chrom_merged$pos, '_', c('start', 'end'))  
54  
55 # 0.4 creation DF  
56 ~ merged_df <- merged_df %>%  
57 ~| mutate(  
58 |   chr = chrom_merged$chr,  
59 |   pos = start_merged$start,  
60 |   X = round(Depth * Meth),  
61 |   X = ifelse(is.na(X), 0, X),  
62 |   N = Depth,  
63 |   N = ifelse(is.na(N), 0, N))
```

Impact de l'âge et du mode d'élevage (cage / sol) sur le méthylome sanguin de la poule.

Stage réalisé dans le cadre du Master 2 BGE, 2025



Réalisé par Jonathan Chalaye

Sous la responsabilité de : Sonia Eynard
Frédérique Pitel

STRUCTURE : INRAE, UMR GENPHYSE (ÉQUIPE GENESIS)

Table des matières

Résumé.....	1
Introduction.....	2
Matériel et méthode.....	5
Protocole expérimental.....	5
Traitement des données.....	7
Analyse différentielle.....	9
Analyse biologique.....	10
Résultats.....	11
Analyse des lectures brutes.....	11
Sorties du pipeline.....	13
Application des filtres.....	15
Caractérisation des données.....	17
Analyse différentielle.....	19
Analyse biologique.....	21
Discussion.....	22
Bibliographie :.....	25
Annexe.....	30
Introduction.....	30
Seuil de filtration de profondeur maximale (Matériel et Méthode).....	30
Benchmarking du seuil maximum (Résultats).....	31
Conclusion du seuil maximum (Discussion).....	33
Barcodes.....	35
Abréviations.....	36
Tableau de DML.....	36

Remerciements

Je tiens à remercier chaleureusement toute l'équipe de **Genphyse** qui m'a accueilli durant ce stage pour son encadrement, sa disponibilité et sa bienveillance, mais aussi l'équipe **GEroNIMO** qui a pu apporter des solutions à mes problèmes. Ce stage m'a permis d'enrichir mes compétences techniques tout en découvrant un environnement de travail stimulant et collaboratif. Je suis particulièrement reconnaissant à mes tutrices de stage, **Sonia Eynard** et **Frédérique Pitel**, pour leurs multiples conseils, leur accompagnement et leur persévérance à mon égard tout au long de cette expérience. Je tiens aussi à remercier **Rémi Séraphin**, le bioinformaticien du projet qui m'a apporté de nombreux conseils au niveau de la pipeline qui est dédiée au projet.

Grâce à ce stage, j'ai pu approfondir mes connaissances, gagner en autonomie et confirmer mon intérêt pour ce domaine.

Contribution personnelle

Durant mon stage, j'ai activement contribué à toutes les étapes du projet GEroNIMO pour l'analyse du méthylome sanguin de la poule en mobilisant les compétences techniques acquises lors de mon master, mais aussi en développant de nouvelles compétences, utiles à la réalisation de ce projet. Voici un résumé de mes principales contributions personnelles :

- Développement de scripts: j'ai réalisé de nombreux scripts en **R**, **Python** et **Bash** permettant de réaliser et d'automatiser les étapes de l'analyse de ce projet.
- Gestion d'un GitHub: j'ai mis en place et maintenu un **dépôt GitHub** pour centraliser l'ensemble des démarches et du code développé durant mon stage. Afin de faciliter le suivi du projet et de faciliter le partage de mes résultats avec mes tutrices lors de réunions hebdomadaires.
- Présentation de travaux:
 - j'ai **présenté les modèles statistiques** utilisés aux collaborateurs du projet GEroNIMO, permettant de recueillir des retours constructifs et d'échanger sur les choix et étapes à adopter.
 - j'ai contribué activement aux **réunions** hebdomadaires avec mes tutrices permettant l'adoption de décisions impactantes l'ensemble du projet.
 - j'ai présenté l'intégralité de mes travaux lors **d'une présentation orale** pendant un séminaire dédié aux stagiaires de l'unité GenPhySE.
- Prise d'initiatives:
 - j'ai proposé un **filtre de seuil maximum personnalisé** pour améliorer le filtrage des données.
 - J'ai également pris l'initiative d'utiliser **Docker** pour obtenir un environnement qui garantit une reproductibilité des résultats et une disponibilité à tous les utilisateurs.
 - Enfin j'ai mis à disposition un **dépôt GitHub en anglais** (<https://github.com/Jonathano3/DSSdiffanalysis>), comprenant des scripts permettant de réaliser l'intégralité de l'analyse ainsi qu'un tutoriel pour réaliser son propre environnement docker.

Contexte du projet

Le projet était déjà défini avant mon arrivée, et toutes les données étaient disponibles sous forme brute. La pipeline de pré-traitement des données était en cours de construction, par **Rémi Séraphin**, lors de mon arrivée. Certains scripts m'ont été transmis par ma tutrice, **Sonia Eynard**, repris d'un projet de thèse antérieur mené par **Chloé Cerutti** ou d'une thèse en cours de **Stacy Rousse**. Cependant, j'ai été pleinement impliqué dans l'adaptation, l'optimisation et l'extension de ces scripts pour les ajuster aux objectifs spécifiques de mon stage, en y apportant également des éléments nouveaux et innovants.

Résumé

Ce projet repose sur l'hypothèse que la méthylation de l'ADN, en tant que mécanisme de régulation génétique, est modulée par l'environnement d'élevage et le vieillissement des poules. En comparant deux types d'élevage (cage et sol) à deux âges (70 et 90 semaines), l'étude cherche à évaluer l'impact de ces facteurs à partir du méthylome sanguin de 1149 individus. Pour cela, ce projet utilise une méthode RRBS (Reduced Representation Bisulfite Sequencing) qui cible les îlots CpG présents sur le génome. Suite à un pipeline utilisant un outil dédié, BISCUIT, et une filtration adéquate de ces données, il en ressort 592 391 sites CpG uniques. Par la suite, une analyse différentielle a été réalisée à partir d'un package R, DSS, se basant sur un modèle multivarié avec interaction de nos deux facteurs, pour détecter des loci différemment méthylés (DML) et des régions différemment méthylées (DMR). Une analyse d'enrichissement (SEA, Singular Enrichment Analysis) a été effectuée sur les DML du facteur d'élevage. On retrouve pour le facteur d'élevage 225 DML hypométhylées, 169 DML hyperméthylées, 9 DMR hypométhylées et 3 DMR hyperméthylées. Dans ces DML, on en trouve 73 qui sont présents dans des gènes et 74 dans des promoteurs. Pour le facteur d'âge, on retrouve 22 DML dont 18 dans des gènes, et pour l'interaction entre ces facteurs, on trouve 2 DML présents dans 2 régions génomiques et 1 dans un promoteur. Les gènes VPS26A et SRA1 sortent significativement affectés par l'élevage en cage ou au sol, ainsi que d'autres gènes qui sont impliqués dans des fonctions biologiques essentielles, telles que le métabolisme, l'immunité, la minéralisation, le système nerveux et rétinien.

Mots clefs : *pipeline, méthylation, analyse différentielle, DML*

This project is based on the hypothesis that DNA methylation, as a mechanism of genetic regulation, is modulated by the rearing environment and the ageing of hens. By comparing two types of rearing (cage and floor) at two ages (70 and 90 weeks), the study seeks to assess the impact of these factors based on the blood methylome of 1,149 individuals. To do this, the project uses a RRBS (Reduced Representation Bisulfite Sequencing) method that targets CpG islands present in the genome. Following a pipeline using a dedicated tool, BISCUIT, and adequate filtering of this data, 592,391 unique CpG sites were identified. Subsequently, a differential analysis was performed using an R package, DSS, based on a multivariate model with interaction between our two factors, to detect differentially methylated loci (DMLs) and differentially methylated regions (DMRs). A Singular Enrichment Analysis (SEA) was performed on the DMLs of the breeding factor. For the breeding factor, there were 225 hypomethylated DMLs, 169 hypermethylated DMLs, 9 hypomethylated DMRs and 3 hypermethylated DMRs. Of these DMLs, 73 are found in genes and 74 in promoters. For the age factor, there were 22 DMLs, 18 of which are in genes, and for the interaction between these factors, there were 2 DMLs present in 2 genomic regions and 1 in a promoter. The VPS26A and SRA1 genes are significantly affected by cage or floor rearing, as are other genes involved in essential biological functions such as metabolism, immunity, mineralisation, and the nervous and retinal systems.

Keywords: *pipeline, methylation, differential analysis, DMLs*

Introduction

Contexte

Dans un contexte de mutations sociétales d'ordre technologique, environnemental et économique, les systèmes d'élevage avicoles font face à des bouleversements majeurs affectant i) un aspect de contrainte de production avec une perpétuelle croissance de la demande alimentaire et un changement climatique majeur, mais aussi ii) un aspect éthique avec une société de plus en plus sensibilisée au bien-être animal, qui était dans les années 1950 mis de côté au profit de la productivité avec l'industrialisation du domaine de l'élevage des poules pondeuses.

En raison d'une augmentation de la croissance démographique mondiale allant de 10 milliards de personnes en 2050 à près de 11 milliards de personnes en 2100 (*United Nations 2017*), la demande en œufs dans les filières avicoles a connu et continue de connaître une augmentation importante. La filière ponte bénéficie d'une empreinte carbone relativement faible, avec 25 g CO₂/g de protéine contrairement aux viandes rouges qui représentent près de 5 à 10 fois cette quantité (*Gaillac R, & Marbach S et al. 2021*) (Figure 1), ce qui s'inscrit dans une ressource plus à l'écoute des demandes de la société qui demande une consommation plus responsable vis-à-vis de l'environnement.

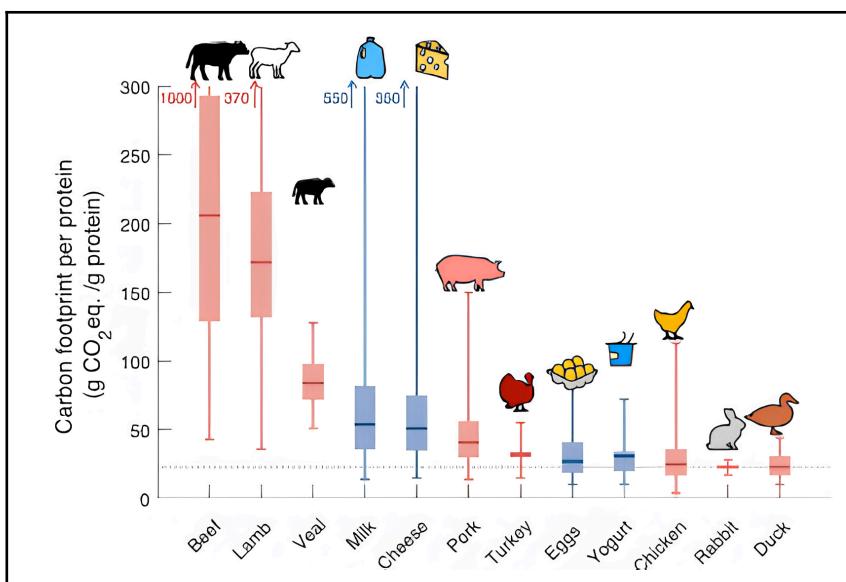


Figure 1: représentation graphique de l'empreinte carbone par g de protéine retenue et calculée pour différents produits carnés et laitiers. La ligne grise en pointillés est une ligne indicatrice correspondant à la valeur la plus faible de l'empreinte carbone par gramme de protéine retenue.

Figure adaptée de : *Gaillac R et al. 2021*.

De plus, la filière ponte joue un rôle crucial grâce à la production de protéines animales par l'exploitation des œufs destinés à la consommation humaine : cette industrie s'inscrit dans un processus d'accessibilité économique abordable pour des petits ménages par l'allocation d'une source importante de protéines, de lipides et de micronutriments sachant que 2 milliards de personnes dans le monde demeurent en manque de micronutriments (*Tulchinsky et al. 2010*).

Il est ainsi prévu que la production mondiale d'œufs atteigne environ 100 millions de tonnes en 2050, ce qui représente une hausse de plus de 30 % dans les prochaines décennies (*Alexandratos et al. 2012*). Cette augmentation significative requiert une optimisation des systèmes de production pour assurer une disponibilité suffisante en œufs tout en limitant l'empreinte écologique de leur production et en satisfaisant les critères de profitabilité économique de la filière et de responsabilité sociétale envers le consommateur.

Une stratégie clé pour répondre à ces besoins consiste à prolonger la durée du cycle de ponte des poules pondeuses au-delà des 60-70 semaines classiques, pour se rapprocher, voire dépasser les 90 semaines tout en gardant la même qualité de produit.

Une augmentation de la période de ponte offrirait de nombreux bénéfices :

- Une augmentation de la production par poule.
- Une optimisation de l'utilisation des ressources alimentaires dédiées au cycle de croissance de la poule.
- Une diminution notable de la taille des lots de poules pondeuses pour la même quantité d'œufs produits, ce qui réduit le nombre de poules retirées à la fin du cycle et restreint le nombre de poussins mâles éliminés après un 1 jour de vie.
- Une baisse de l'empreinte écologique associée à l'élevage de moins de poules pondeuses.
- Un amoindrissement pour les éleveurs des coûts finaux de production.

Parallèlement, le choix du mode d'élevage est un enjeu majeur pour le consommateur, qui est de plus en plus soucieux des besoins physiologiques et comportementaux des animaux qu'il consomme. En effet, il est attendu que les poules au sol manifestent davantage de comportements normaux que leurs congénères en cage qui sont moins actifs et plus stressés avec des comportements plus agressifs avant la ponte (*Urban-Chmiel, R. 2010*). Tandis que des comportements plus naturels pour des poules au sol tels que picorer et gratter réduisent le risque d'interactions sociales négatives et contribuent ainsi à leur bien-être. Le débat opposant les systèmes d'élevage en cage et au sol est ainsi au cœur des préoccupations sociétales actuelles, notamment grâce à l'initiative citoyenne européenne (ICE) "End the Cage Age" (*European Commission. 2021*) signée par 1,4 million de citoyens visant à l'interdiction de l'utilisation de la cage pour un certain nombre d'animaux d'élevage (poules pondeuses, lapins, poulets de chair en reproduction, etc...). Dans cette perspective, pour assurer la pérennité et la durabilité de la production d'œufs, il est essentiel d'intégrer des innovations techniques et des approches scientifiques collaboratives entre les industries avicoles et la communauté scientifique.

Objet de l'étude

Les systèmes d'élevage doivent s'adapter aux changements climatiques ainsi qu'aux évolutions éthiques actuelles, qui transforment en profondeur les attentes sociétales et les modes de production. Dans ce cadre, l'adaptation des poules pondeuses est un facteur majeur pour assurer la durabilité de ce système de production. Cette adaptation repose d'une part sur un aspect transmis entre générations par la sélection génétique de lignées optimales, mais aussi sur une plasticité, c'est-à-dire la capacité d'un même génome à répondre de manières différentes en fonction de stimuli extérieurs. L'un des moteurs de cette plasticité et de cette adaptation est constitué par des mécanismes épigénétiques, qui agissent à l'interface entre le génome et l'environnement. L'épigénétique regroupe un ensemble de modifications moléculaires qui régulent l'expression des

gènes sans altérer la séquence de l'ADN. Parmi ces phénomènes on retrouve les modifications des histones, les ARN non-codants régulateurs d'expression génique, mais aussi la méthylation de l'ADN. C'est sur cette dernière modification que nous allons nous pencher dans le cadre de notre étude.

Dans le cadre de notre étude nous allons nous intéresser à la méthylation de l'ADN, sur les motifs CpG, majorité des sites méthylos chez les vertébrés. Ces dinucléotides peuvent en effet être modifiés par l'ajout d'un groupement méthyle qui va se fixer sur la cytosine, on appelle ce processus la méthylation. Cette régulation épigénétique chez la poule présente une particularité, le génome présente un niveau global d'hypométhylation plus élevé que chez la plupart des autres vertébrés (*Al Adhami et al., 2022*). La méthylation est largement connue pour contribuer à la régulation de l'expression génique. Le processus de méthylation affecte la structure 3D des molécules d'ADN influençant ainsi la stabilité du génome, entraînant l'inactivation de certains éléments transposables et la régulation fine de nombreux processus biologiques, comme le développement cellulaire et la réponse aux stimuli environnementaux.

L'impact le plus direct des méthylations de l'ADN reste leur influence sur l'expression génique, qui par leur présence dans les régions promotrices des gènes notamment, peuvent empêcher physiquement la liaison entre les facteurs de transcription et l'ADN de se former et bloquer ainsi l'initiation de la transcription des gènes en amont par l'ARN polymérase. Ce processus est un véritable outil de régulation de l'expression génique car il constitue un "interrupteur moléculaire" qui permet de moduler l'activité des gènes en fonction des stimuli et des besoins de la cellule. Ainsi, en général, un gène dont le promoteur est très méthylos est moins exprimé qu'un gène dont le promoteur est peu méthylos. De plus, les méthylations présentes dans les exons et les introns peuvent influencer l'épissage de l'ARN pré-messager en empêchant la reconnaissance des sites dédiés par les facteurs d'épissage ou altérer la structure locale de la chromatine. Ces méthylations permettent à certains exons d'être inclus ou exclus de l'ARNm contribuant à la diversité des isoformes produites à partir d'un même gène et jouent donc un rôle essentiel dans la régulation post-transcriptionnelle de l'expression génique.

La méthylation de l'ADN s'inscrit aussi dans un rôle plus large en participant à la conservation des régions essentielles avec une méthylation plus faible des régions très bien conservées (*Li et al., 2011*) reflétant un marqueur épigénétique incroyablement stable contrairement à l'expression des gènes qui peut varier rapidement en réponse à des conditions variables.

Ce stage vise à analyser l'impact de l'allongement du cycle de la période de ponte et du système d'élevage (sol/cage) sur le méthylome de la poule. L'analyse de marques épigénétiques dans le méthylome de la poule en condition cage ou sol et pour des poules ayant 70 semaines ou 90 semaines permettra de mieux comprendre les mécanismes d'adaptation à ces conditions de stress et leurs conséquences potentielles sur des mécanismes biologiques associés.

L'hypothèse de cette étude est que l'architecture de régulation génétique est soumise à l'environnement et évolue avec le temps, ce qui est le cas chez l'homme et la souris (*Feil, 2006*) (*Romero et al., 2012*). Ces mécanismes étant moins connus chez les oiseaux, on sait tout de même que la méthylation est un marqueur connu de l'âge qui décroît avec les années (*Bollati et al., 2009*; *Gryzinska et al., 2013*) et que la méthylation a un lien avec la régulation des gènes qui conduit alors à un changement d'expression des gènes au cours de la vie des individus (*Attwood et al., 2002*). Par conséquent, examiner l'impact de l'allongement de la carrière des poules pondeuses en comparant 2 environnements (cage/sol) à deux âges (70/90 semaines) permettra de mieux

comprendre comment l'adaptation aux conditions d'élevage et le vieillissement influencent la méthylation dans les cellules sanguines de cette espèce.

L'objectif est d'identifier les marques épigénétiques, en particulier la méthylation de l'ADN, affectées par l'allongement du cycle de ponte des poules pondeuses ou par le type d'élevage. L'analyse des motifs CpG et de leurs modifications dans les cellules sanguines selon les conditions d'élevage et l'âge des poules pondeuses permettra de déceler des indicateurs épigénétiques de méthylation associés à l'adaptation des individus aux conditions d'élevage. L'étude des profils de méthylation contribuera à une meilleure compréhension des mécanismes biologiques en lien avec la productivité, la longévité et le bien-être des poules.

De plus, cette étude contribuera à l'enrichissement des connaissances scientifiques sur la méthylation de l'ADN dans le contexte de l'élevage avicole, en préparant le terrain pour des études supplémentaires sur l'adaptation génétique et épigénétique à des modifications de l'environnement et aux potentielles méthodes d'amélioration des techniques d'élevage.

Matériel et méthode

Protocole expérimental

Notre étude comprend 691 poules, issues d'une lignée pure de poules pondeuses provenant de l'élevage du partenaire privé NOVOGEN. Cette lignée a été sélectionnée à partir de poules de la race Rhode Island Red à œufs bruns, fréquemment utilisée dans les programmes de sélection du fait de leur période de ponte importante et de leur bonne qualité des œufs.

Élevées dans des parcs au sol jusqu'à 17 semaines, 480 poules ont ensuite été transférées en cages collectives (cinq poules par cage), tandis que 211 ont été placées en volière. À 55 semaines, tous les animaux ont été installés en cages individuelles jusqu'à leur abattage à 93 semaines pour 457 d'entre elles (les autres ayant été abattues à 70 semaines pour une autre étude). Les poules ont été nourries "ad libitum" avec un aliment de ponte commercial (11,47 MJ/kg d'énergie métabolisable, 16,5 % de protéines brutes, 3,70 % de calcium), sous un régime lumineux de 16 h de lumière et 8 h d'obscurité, à une température constante de 20 °C (*Berger et al., 2025*).

Des prélèvements sanguins ont été effectués sur l'ensemble des animaux immédiatement après leur transfert en cage individuelle à 70 semaines puis sur 457 poules à 90 semaines. L'ADN a été extrait à partir des globules rouges (*Roussot et al., 2003*) de ces échantillons de sang.

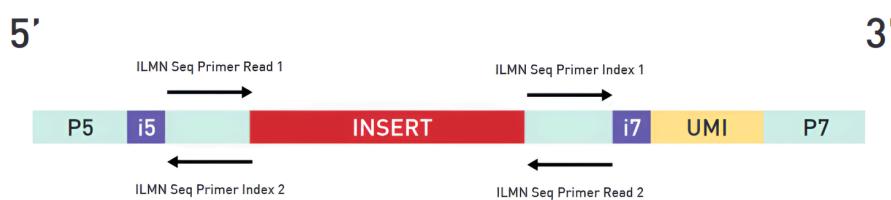
La méthode Reduced Representation Bisulfite Sequencing (RRBS) a été utilisée pour les analyses de méthylation. C'est une stratégie ciblée pour examiner le profil de méthylation de l'ADN sur une portion du génome. À la différence du Whole Genome Bisulfite Sequencing (WGBS), qui consiste à séquencer la totalité du génome soumis au bisulfite, le RRBS se focalise spécifiquement sur les zones enrichies en dinucléotides CpG, comme les îlots CpG ou plus particulièrement les sites promoteurs qui sont particulièrement impactées par la méthylation. La méthode RRBS (*Nakabayashi et al., 2023*) est obtenue suite à une digestion enzymatique initiale de l'ADN à l'aide de l'enzyme MspI, qui cible les sites C/CGG, suivie d'une sélection de fragments d'ADN selon leur

taille afin d'enrichir les échantillons en régions assez informatives (si trop petit pas assez de CpG ou artefact et si trop grand généralement région pauvre en CpG et trop coûteux pour le séquençage). Les fragments sélectionnés sont ensuite soumis à un traitement au bisulfite de sodium, qui convertit les cytosines non méthylées en uraciles (puis en thymines après amplification), tandis que les cytosines méthylées demeurent inchangées. Le séquençage de ces fragments permet ainsi d'identifier précisément les sites de méthylation.

Les données de séquençage RRBS ont été produites à l'aide du kit “premium RRBS V2” de Diagenode de préparation de librairies, réalisé au sein du laboratoire, qui a pour but d'optimiser la répétabilité de la méthode sur l'intégralité de nos échantillons. Le séquençage a ensuite été confié à Azenta.

Le protocole Diagenode s'appuie sur une séquence normalisée d'étapes pour assurer une analyse RRBS précise et répliable. Initialement, l'ADN récupéré subit une digestion enzymatique avec l'enzyme MspI. Les fragments obtenus sont par la suite préparés pour la ligation, ce qui permet l'intégration d'adaptateurs Illumina et d'identifiants moléculaires uniques (UMI), indispensables pour minimiser les biais de la PCR et garantir une mesure précise de la méthylation. Un processus de sélection de taille des fragments à séquencer, suivi d'une quantification individuelle et d'un mélange (“pooling”) des échantillons, précède l'étape de conversion au bisulfite. Cette dernière transforme les cytosines non méthylées tout en conservant celles qui sont méthylées, l'ADN transformé est par la suite amplifié par PCR avant le séquençage.

Ensuite, l'ADN traité est séquencé par Azenta à l'aide de la technologie Illumina, qui permet de générer des lectures de haute précision avec un débit de séquençage cible de 10 millions de lectures par échantillon.



P5/P7 = Illumina adapters

UMI = Unique Molecular Identifier

i5/i7 = Unique Dual Indexes

Figure 2: Système de construction des échantillons séquencés après traitement par le premium RRBS kit V2 de Diagenode. La construction finale porte des primers Illumina, des Unique Dual Indexes (UDI), des identifiants moléculaires uniques (UMI) et des adaptateurs Illumina. RRBS

Traitement des données

L'analyse des données obtenues avec ce séquençage RRBS a été réalisée par le biais d'une pipeline inspirée de la pipeline nf-core methylseq, qui utilise les outils Bismark ou bwa-meth. A terme, cette pipeline a pour vocation d'être intégrée dans la pipeline nf-core methylseq pour ajouter un outil supplémentaire traitant les données RRBS : BISCUIT (*Zhou et al. 2024*).

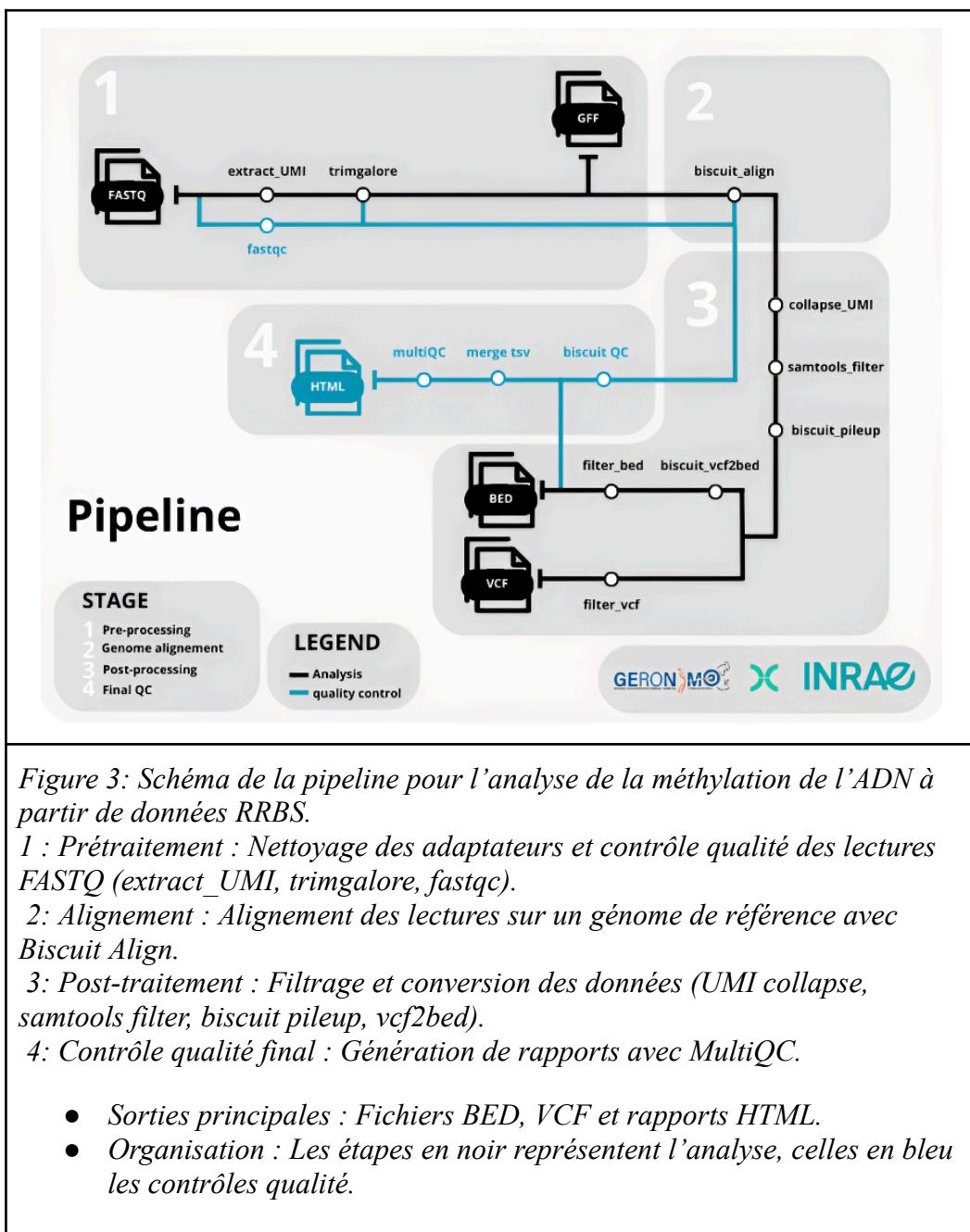


Figure 3: Schéma de la pipeline pour l'analyse de la méthylation de l'ADN à partir de données RRBS.

1 : Prétraitement : Nettoyage des adaptateurs et contrôle qualité des lectures FASTQ (extract_UMI, trimgalore, fastqc).

2: Alignement : Alignement des lectures sur un génome de référence avec Biscuit Align.

3: Post-traitemet : Filtrage et conversion des données (UMI collapse, samtools filter, biscuit pileup, vcf2bed).

4: Contrôle qualité final : Génération de rapports avec MultiQC.

- Sorties principales : Fichiers BED, VCF et rapports HTML.
- Organisation : Les étapes en noir représentent l'analyse, celles en bleu les contrôles qualité.

La pipeline (Figure 3) a été spécialement conçue pour analyser les données de méthylation issues du séquençage au bisulfite. Elle repose principalement sur l'outil BISCUIT, qui est un outil d'alignement spécialisé dans ce type de données. BISCUIT présente l'avantage d'aligner efficacement les lectures traitées au bisulfite sur des génomes de référence, mais aussi d'intégrer des fonctionnalités uniques par rapport à d'autres outils d'alignement : il permet l'identification de variants et la gestion des trous (“gaps”) dans les alignements, ce qui facilite une évaluation plus précise des profils de méthylation, tout en considérant la variabilité génétique et les éventuelles anomalies dans les séquences.

La pipeline se déroule ainsi: 1/ tout d'abord, un contrôle de qualité est réalisé sur les séquences fasta avec **fastqc** (*Andrews S., 2010*), puis en parallèle, l'étape **Extract_UMI** utilise **umi-tools** (*Smith et al., 2017*) afin d'extraire les UMIs présents dans les séquences. Les UMIs permettent d'identifier chaque molécule d'ADN de façon unique. Cela aide à distinguer les séquences provenant de molécules différentes, même si elles sont identiques, et donc à éviter de compter plusieurs fois les mêmes lectures dues à l'amplification PCR.

Ensuite, **Trimgalore** (*Martin, 2011*) est utilisé pour supprimer les adaptateurs R1 “AGATCGGAAGAGCACACGTCTGAACCTCCAGTCA” et R2 “AGATCGGAAGAGCGTCGTAGGGAAAGAGTGT” et supprimer les 2 premières et dernières nucléotides de chaque séquence trimmée avec l'option *-clip_R1 2 -clip_R2 2 -three_prime_clip_R1 2 -three_prime_clip_R2 2* ce qui permet de supprimer les sites de digestion par l'enzyme Mspl.

La seconde étape de la pipeline est l'alignement des lectures sur le génome de la poule (version de l'assemblage bGalGal1.mat.broiler.GRCg7b d'ID NCBI GCF_016699485.2). Cet alignement est réalisé à l'aide de **Biscuit Align**, qui prend en compte les conversions des cytosines en uraciles en convertissant *in silico* le génome de référence comme s'il avait subi un traitement au bisulfite, pour améliorer le taux d'alignement. Une fois les séquences alignées, des étapes de post-traitement 3/ et de contrôle de qualité 4/ avec **biscuitqc** nous permettent de contrôler la qualité des alignements ainsi que de s'assurer que les adaptateurs ont bien été retirés.

A posteriori, l'étape **Collapse_UMI** permet de regrouper les lectures issues d'un même fragment d'ADN à partir de leur identifiant UMI afin d'éliminer les séquences dupliquées. Par la suite, un filtre est appliqué à l'aide de **Samtools_filter** (*Danecek et al., 2021*) pour éliminer les alignements de qualité inférieure à 40 (option *-q*) et les lectures secondaires non informatives s'alignant à plusieurs endroits dans le génome, qui sont marquées avec un flag 256 (option *-F* de l'outil).

L'étape **Biscuit_Pileup** permet ensuite de générer un fichier vcf qui va identifier les variants. Ensuite, **Biscuit_vcf2bed** convertit les données de méthylation en un format BED, adapté aux analyses : pour chaque position, le logiciel compte le nombre de lectures comprenant un C (méthylé) ou un T (non méthylé) et calcule le taux de méthylation à cette position par le rapport C/(C+T). Enfin, les étapes **Filter_bed** et **Filter_vcf** vont appliquer des filtres pour obtenir une profondeur minimale de 10 et éliminer un millième des positions les plus profondes, pour que les ratios de méthylation possèdent une certaine robustesse quant à leurs valeurs. Pour finir, **merged_tsv** est une étape du contrôle de qualité 4/ qui rassemble avec l'aide d'un **multiqc** (*Ewels et al., 2016*) certaines statistiques obtenues tout au long de la pipeline (fastqc, biscuitqc) sur toutes les étapes et échantillons dans un fichier HTML.

Ce pipeline nextflow permet ainsi d'obtenir les sorties BED, VCF et un rapport HTML condensé comprenant les différents contrôles de qualité.

À l'issue de ce pipeline, un total de **2 190 123 sites CpG uniques** a pu être identifié à travers l'ensemble des échantillons analysés. Ces données constituent la base de l'analyse.

Analyse différentielle

L'analyse différentielle consiste à identifier des sites CpG dont le taux de méthylation varie significativement entre deux ou plusieurs conditions. Elle repose sur des tests statistiques appliqués à chaque élément pour détecter des différences liées à un facteur d'intérêt, élevage et période dans notre cas.

L'analyse différentielle de méthylation a été réalisée en langage R (version 4.4.0) et à l'aide du package Dispersion Shrinkage for Sequencing data (DSS) (version 2.54.0) (*Feng et al., 2014; Wu et al., 2013*), permettant le traitement statistique des données de méthylation issues du séquençage au bisulfite. Cet outil d'analyse a comme spécificité de pouvoir attribuer des poids pour chaque valeur de méthylation sur les CpG en fonction de la profondeur de lecture, ce qui permet de pouvoir prendre en compte toutes les valeurs de méthylation, y compris celles de faible couverture. DSS utilise différents types de modèles afin de détecter les loci différentiellement méthylés (DML) entre plusieurs conditions :

- Un modèle linéaire généralisé (GLM) simple qui suppose une distribution bêta-binomiale, qui est particulièrement adaptée au traitement des données de comptage. Il s'utilise pour comparer deux groupes et donc une condition binaire. Ce modèle estime la moyenne de méthylation pour chaque site CpG puis leur dispersion afin de conduire un test de Wald qui va permettre de distinguer les CpG différentiellement méthylés pour la condition testée. Le test différentiel est réalisé avec la fonction DMLtest().
- Un même modèle linéaire généralisé multifacteur est mis à disposition dans DSS pour les dispositifs complexes, comprenant plusieurs variables explicatives. Ce modèle multifactoriel offre la possibilité de maîtriser l'impact des variables confondantes tout en mettant en évidence l'effet principal d'un facteur d'intérêt sur la méthylation. Pour ajuster les données de méthylation à ce GLM, DSS utilise une transformation arcsinus, spécifiquement sélectionnée pour les données de méthylation. Cette conversion vise à stabiliser la variance aux bornes de l'intervalles de méthylation (0 et 1), où la variance est particulièrement basse. Par la suite, le modèle ajuste les données transformées en utilisant la technique des moindres carrés et réalise un test de F-test afin d'identifier les sites CpG où des variations significatives de méthylation sont associées aux facteurs examinés. L'ajustement est réalisé via la fonction DMLfit.multiFactor(), et les tests différentiels à l'aide de DMLtest.multiFactor().

De plus, pour tous ces modèles de DSS, une option a été implémentée permettant d'effectuer un processus de lissage (« smoothing ») sur les taux de méthylation à travers les sites CpG voisins. Cette méthode se base sur l'hypothèse que la méthylation fluctue de façon plutôt continue à travers le génome, notamment dans les zones fonctionnelles telles que les promoteurs ou les îlots CpG. Le lissage permet ainsi de minimiser le “bruit technique” associé aux couvertures faibles et à la variabilité locale, en fusionnant les données des CpG adjacents pour générer une estimation plus solide et significative sur le plan biologique du profil de méthylation.

Du fait de la structure particulière de nos données qui comprennent 708 échantillons appariés par le facteur semaine et 321 qui ne le sont pas, deux approches distinctes ont été mises au point afin de rendre compte de ces contraintes techniques dues à l'appariement d'un certain nombre de données. En effet, statistiquement il n'est pas trivial de prendre en compte dans le même modèle des données appariées et des données non-appariées et aussi d'intégrer un maximum d'informations. Les modèles sont les suivants, avec **cage** qui représente la condition cage/sol et **w** qui représente la période 70/90 semaines :

- **Modèle simple:** ~ **cage** : Ce modèle a été construit à partir de deux sous-ensembles d'échantillons définis par le facteur "semaine", qui contient des données appariées. Il inclut 597 échantillons à la semaine 70 et 432 échantillons à la semaine 90. Le modèle DSS utilisé est un modèle GLM simple smoothed, permettant d'évaluer uniquement l'effet de la condition "cage". L'objectif de ce modèle est de contourner les contraintes techniques liées à l'appariement.
- **Modèle multifacteur :** ~ **w + cage + cage:w** : Ce modèle a été construit seulement sur les échantillons appariés (708 échantillons), il comprend un modèle GLM multifactoriel smoothed plus complexe qui a été utilisé pour étudier les interactions "cage" et "semaine", l'effet de la condition "semaine" et de la condition "cage". L'objectif de ce modèle, bien que contraint par l'appariement des données, est d'étudier tous les effets des facteurs et aussi les effets combinés des deux facteurs sur les données de méthylation.

À la suite de l'ajustement des différents modèles statistiques, une analyse des régions différentiellement méthylées (DMR) a été réalisée sur nos DML à l'aide de la fonction callDMR(). Cette étape permet de regrouper les sites CpG proches ayant des profils de méthylation similaires et étant identifiés comme des DML significativement différents. Les DMR sont définies selon plusieurs critères :

- une distance maximale entre sites
- une taille minimale de la région
- un nombre minimal de sites significatifs par région
- une p-valeur ajustée minimale

Les seuils suivants ont été appliqués : une distance maximale de 50 paires de bases, une taille minimale de 50 paires de bases, au moins 3 sites significatifs par région, et un seuil de p-valeur ajustée d'au moins 1 %. Cette approche permet de renforcer la robustesse biologique des résultats, en se concentrant sur des segments entiers du génome plutôt que sur des CpG isolés, et d'interpréter plus facilement les impacts potentiels des méthylations sur la régulation génique.

Analyse biologique

Afin de mieux comprendre les implications fonctionnelles des gènes associés aux régions différentiellement méthylées, une analyse standard d'enrichissement (SEA) a été réalisée. Cette approche permet d'identifier des voies biologiques, processus moléculaires ou fonctions cellulaires

significativement représentés parmi une liste de gènes d'intérêt. Dans notre cas, une liste personnalisée de gènes a été utilisée comme gènes d'intérêt, car notre analyse RRBS est une sous-représentation du génome de la poule, l'analyse ne peut donc être faite que sous cette sous-représentation du génome. Par conséquent, notre base de référence a été définie comme l'ensemble des gènes contenant au moins un site CpG couvert par notre analyse, afin de garantir une comparaison statistique cohérente dans le test d'enrichissement.

Cette liste de gènes est ensuite utilisée pour vérifier si une fonction des gènes de notre liste “custom” présente une surreprésentation spécifique de cette fonction, en adoptant une approche statistique qui suppose l'indépendance des gènes de type Over-Representation Analysis (ORA).

Pour effectuer cette analyse SEA, le package clusterProfiler (version 4.14.6) (Yu et al., 2012) a été utilisé sous R ainsi que la version 113 d'ensembl. Cet outil puissant offre une implémentation robuste de la méthode SEA via la commande enricher() qui nous permet de renseigner notre propre liste de gènes de référence. Il permet également une visualisation claire (dotplot, networks..) et interactive des résultats, facilitant leur interprétation. Par ailleurs, GEGA a été utilisée pour trouver des informations sur le génome de la poule (*Degalez et al., 2024*).

Résultats

Analyse des lectures brutes

La distribution des données brutes RRBS révèle une dispersion homogène entre les échantillons avec une valeur médiane de 10 620 370 lectures, qui est légèrement supérieur au nombre de lectures moyennes attendu de 10 000 000. Néanmoins, certaines anomalies demandent une considération spécifique. Effectivement, deux échantillons (ID 888125 et 871202) se démarquent par une quantité de lectures remarquablement élevée, pouvant aller de 5 à 10 fois plus que la quantité attendue. Par ailleurs, trois autres échantillons (ID 871146, 881212, 887837) affichent un nombre de lectures anormalement bas, fluctuant entre 1 000 et 55 000 reads. De plus, la plaque 10 se distingue par deux clusters distincts dans la distribution de ces reads avec des échantillons qui présentent des reads proches de 0 et d'autre légèrement supérieurs à la distribution globale des données, ce qui introduit une variance inattendue pour cette plaque (figure 4). En fouillant plus en profondeur la répartition des lectures entre les échantillons, on ne remarque pas de barcodes présentant une augmentation ou une diminution notable du nombre de lectures. La distribution semble globalement homogène, ce qui suggère une répartition équilibrée du séquençage entre les différents barcodes (figure 5).

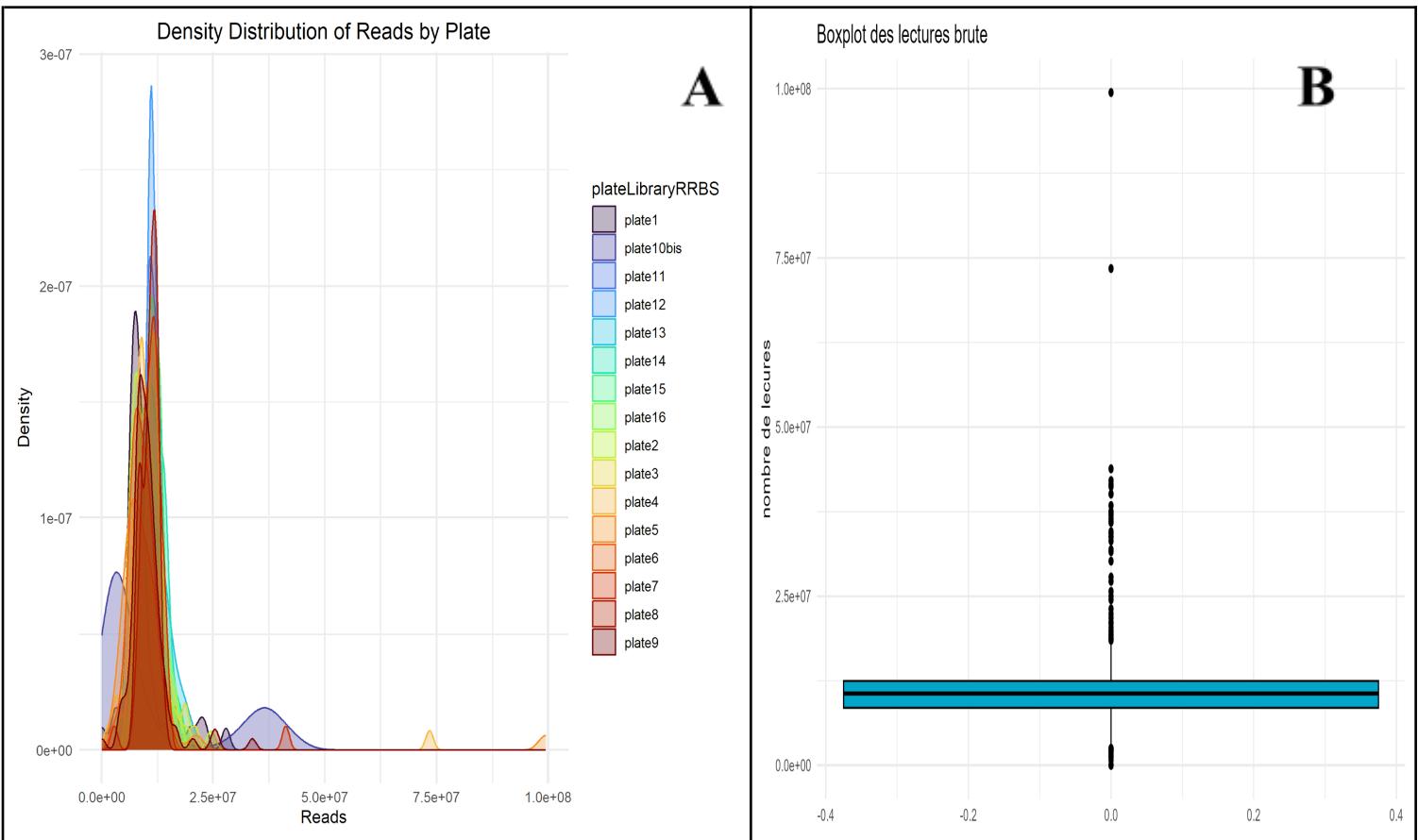


Figure 4: Visualisation de la distribution des lectures entre échantillons et par plaque.

Le graphique A représente la distribution des lectures entre les différentes plaques et le graphique B représente la distribution de l'ensemble des lectures.

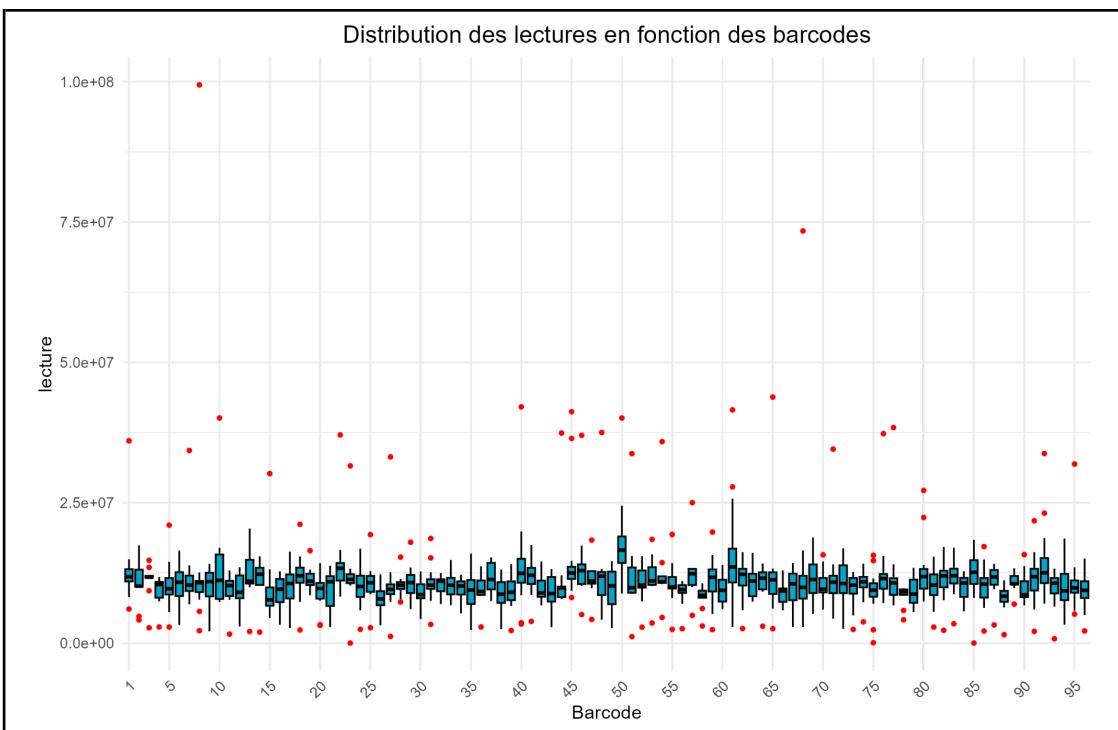


Figure 5: visualisation de la distribution du nombre de lectures des échantillons au sein des différents barcodes utilisés. Les barcodes sont numérotés de 1 à 96 et leur correspondance est détaillée en annexe.

Sorties du pipeline

Suite à l'utilisation du pipeline, un ensemble conséquent de statistiques et de représentations graphiques a été généré pour l'ensemble des échantillons traités et compilés dans un rapport MultiQC. Différentes étapes de contrôle ont été réalisées, notamment le trimming des adaptateurs, à la fois pour les séquences Illumina et les UMIs. Le contrôle qualité de ce trimming indique que la suppression des séquences d'adaptateurs a été correctement effectuée pour l'intégralité des lectures des échantillons, avec un score de qualité Phred qui avoisine les 40. De plus, en regardant la constitution des lectures en adaptateurs le long de la séquence, on constate une augmentation linéaire d'allure extrêmement homogène entre les échantillons, allant de 3% à 10% des séquences qui possèdent un morceau de séquence d'adaptateur jusqu'à 140 pb, ce qui montre une absence significative de contamination résiduelle par les adaptateurs (figure 6). Néanmoins, deux échantillons, B274 et B539, se démarquent du reste du jeu de données par la présence persistante de séquences d'adaptateurs. En regardant plus en détail la composition de ces échantillons anormaux, on constate que le nombre de lectures R1 et R2 est très faible pour ces deux individus : 55 154 lectures pour B274 et seulement 3 056 pour B539. Ces deux échantillons ne suivent pas les profils habituels observés dans les figures de qualité de base du rapport MultiQC.

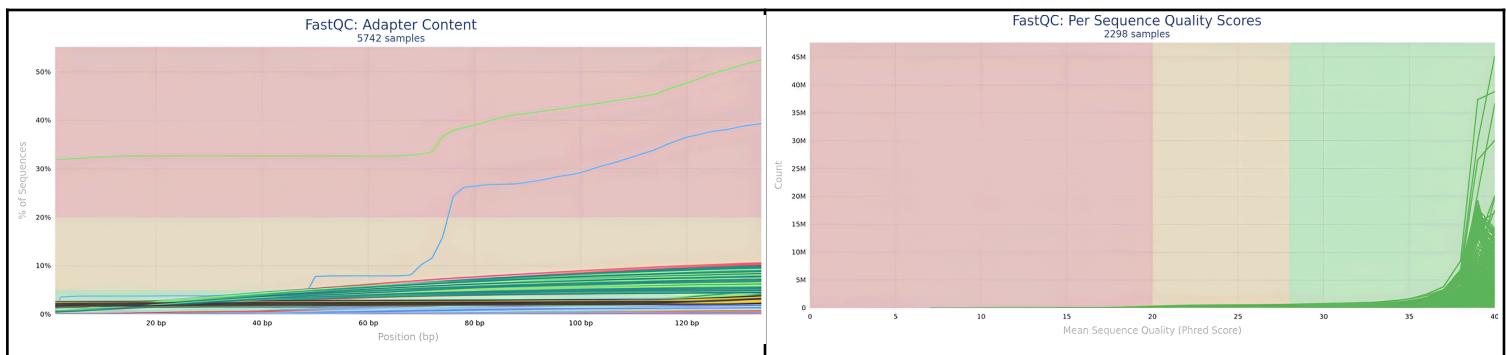


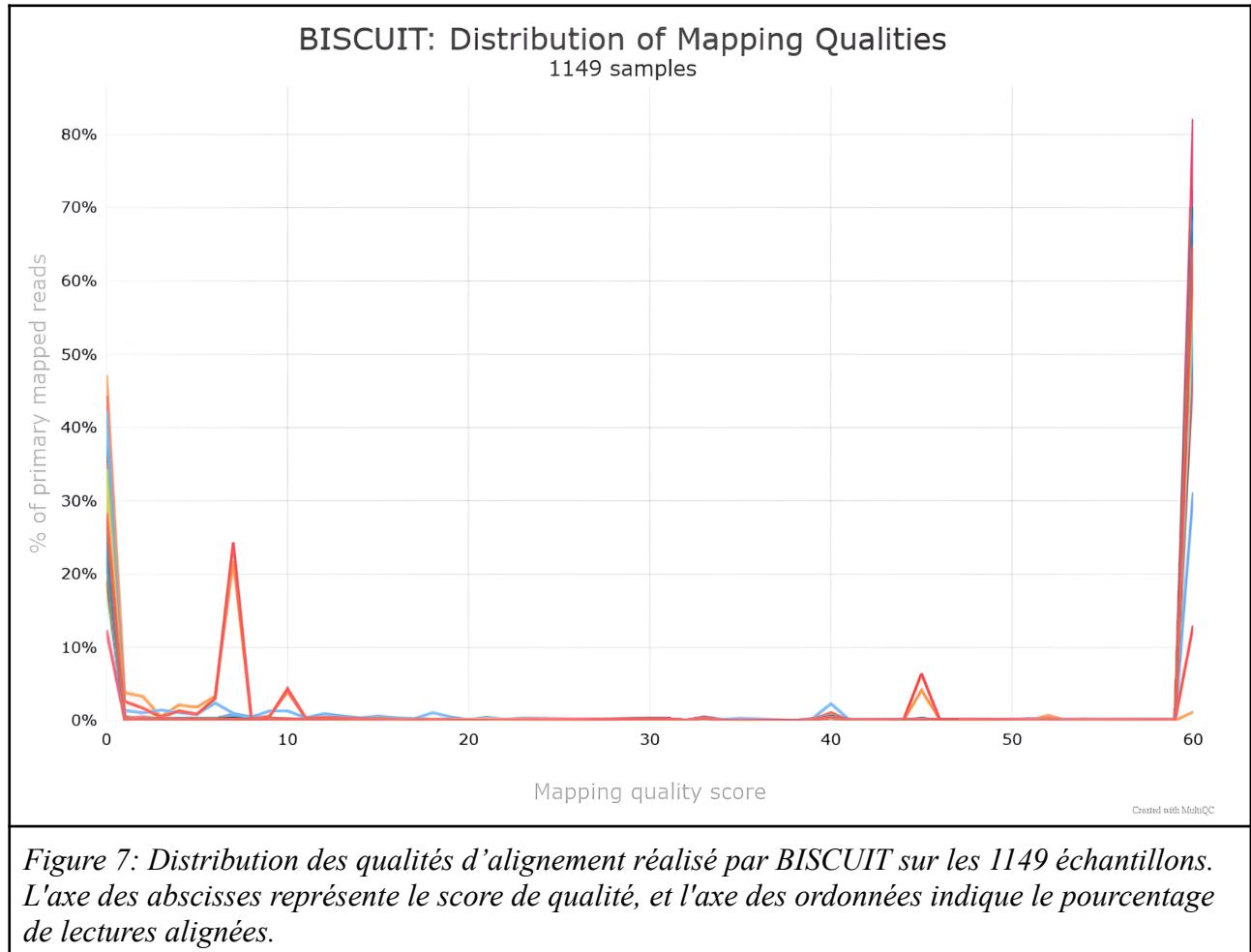
Figure 6: Sortie multiqc de l'évaluation de la qualité des lectures par FastQC.

Le graphique de gauche réfère au contenu en adaptateurs par position dans les lectures : il présente la proportion de séquences identifiées comme dérivées d'adaptateurs à chaque position le long des lectures.

Le graphique de droite réfère au score Phred de qualité moyenne par séquence sur l'ensemble des lectures des échantillons.

La figure 7 présente un contrôle d'une autre étape clé de la pipeline, suivant le mapping réalisé par l'aligner BISCUIT. Ce contrôle porte sur l'évaluation de la qualité d'alignement des lectures. La distribution des scores d'alignement met en évidence deux principaux pics de valeurs. Le premier correspond à des lectures faiblement alignées, avec un score de qualité de 0 et pouvant concerner jusqu'à 50 % des lectures pour certains échantillons. Le second pic traduit une qualité d'alignement élevée, avec des scores proches de 60, et représente jusqu'à 80 % des lectures dans plusieurs échantillons. On retrouve toujours les mêmes échantillons qui posaient problème à l'étape précédente et qui ici se distinguent par une grande proportion de score de faible qualité.

Par ailleurs, des statistiques sur les profondeurs ont été recueillies et confirment l'application correcte du seuil **minimal de 10** défini dans la pipeline. En ce qui concerne la profondeur maximale, elle varie selon les individus et peut atteindre jusqu'à **992 lectures** dans certains cas. On dénombre **7 826 168 CpG unique** avant les filtres de profondeur. Puis **2 190 123 CpG uniques** suite à l'application du filtre, pour une moyenne de **3 652 151 sites couverts par échantillon**, détectés à la suite de ce pipeline.



Application des filtres

Afin de réaliser un traitement sur les sites CpG, un processus rigoureux de filtration a été appliqué afin de garantir la qualité et la fiabilité des données de méthylation issues du séquençage RRBS.

Tout d'abord, deux filtres ont été mis en place directement dans la pipeline, avec un **filtre de profondeur minimum** qui a été ajusté à 10 et un **filtre de profondeur maximum**, avec seuls les sites CpG présentant une couverture inférieure au seuil de 1/1000 de la distribution globale des profondeurs qui ont été conservés, ce qui nous fait perdre **5 636 045 sites CpG uniques** dès **7 826 168 CpG uniques post-pipeline**.

Ensuite, une **filtration sur la qualité** des échantillons est nécessaire afin d'enlever les échantillons qui ne présentent pas suffisamment de sites CpG (figure 8). Un seuil de 200 000 sites CpG minimum a été choisi, nous faisant perdre **120 échantillons** et **862 sites CpG uniques** communs à ces échantillons perdus (figure 8).

Par la suite, une sélection basée sur le polymorphisme a été mise en œuvre dans le but d'exclure les sites susceptibles d'être affectés par des variants génomiques. Ainsi, nous avons combiné les variants identifiés directement à partir des alignements au bisulfite suite à la sortie de variants de l'outil BISCUIT, qui est spécifique à ce logiciel, avec ceux provenant d'un calling réalisé sur ces mêmes lignées sur des données de DNAseq, par nos collaborateurs de l'institut agro de Rennes. Les sites CpG co-localisant avec un variant identifié dans l'une ou l'autre référence ont été exclus de l'analyse, pour éviter toute confusion d'interprétation entre conversion au bisulfite et polymorphisme C/T. Ce **filtre variant** nous a fait perdre **101 403 sites CpG uniques**.

Enfin, un **filtre sur les données manquantes** a été appliqué afin d'assurer la robustesse statistique de l'analyse. Les sites CpG pour lesquels plus de 25 % des échantillons présentaient une absence de couverture ont été exclus du jeu de données final. Nous avons éliminé ainsi **1 495 467 sites CpG uniques**, conduisant à un total de **592 391 sites CpG uniques**, compris dans les **12 749 701 sites CpG uniques** du génome de la poule, qui vont être utilisés pour l'analyse différentielle, avec une couverture globale moyenne de **22.78 X** (figure 9).

Visualisation du seuil choisi pour le filtre de qualité

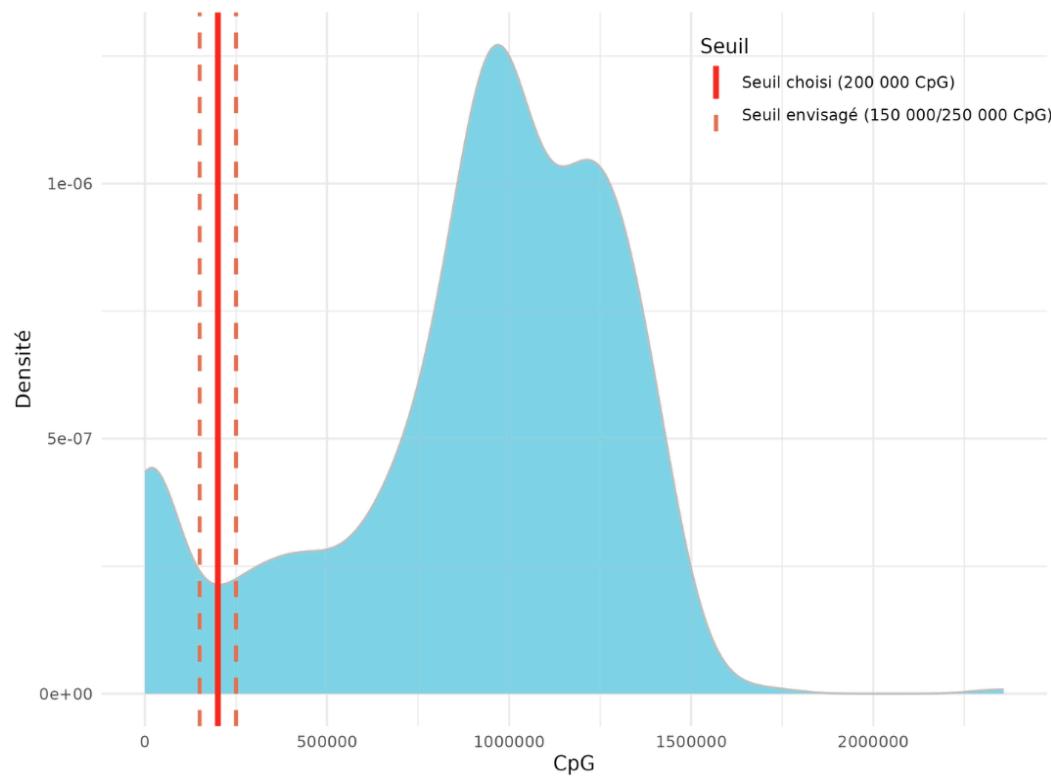


Figure 8: Visualisation du seuil choisi pour le filtre de qualité. La ligne verticale rouge indique le seuil retenu (200 000 CpG), tandis que les lignes pointillées rouges représentent les seuils envisagés (150 000 et 250 000 CpG). La densité des données est affichée sur l'axe des ordonnées, et l'axe des abscisses représente le nombre de CpG. Cette analyse permet de déterminer le seuil optimal pour éliminer des échantillons de faible qualité tout en conservant un volume suffisant d'échantillons pour l'étude.

Visualisation de la perte des sites CpG par les différents filtres

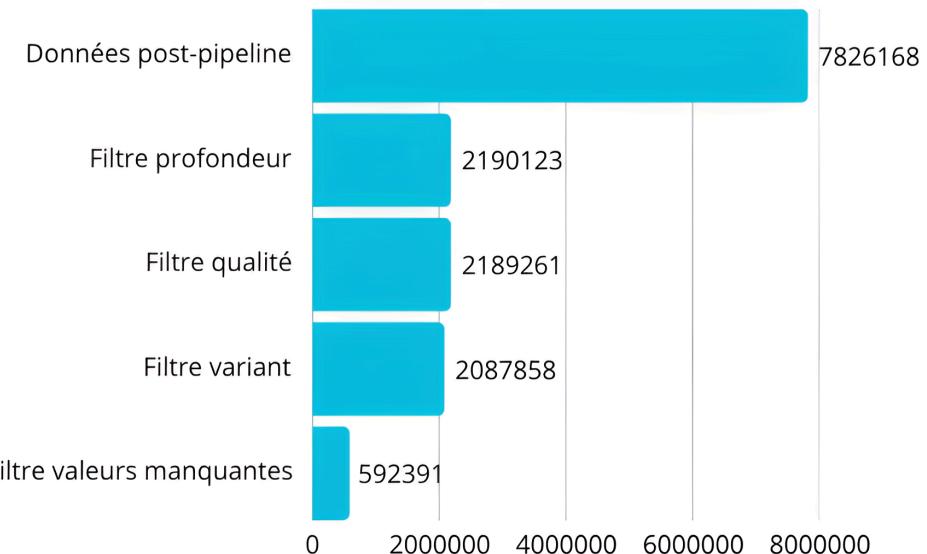


Figure 9: Visualisation de la perte progressive des sites CpG suite à l'application successive de différents filtres sur les données après qu'elles soient passées par le pipeline. Le graphique présente le nombre de sites éliminés à posteriori de chaque étape : filtre de profondeur (qui est intégré dans la pipeline), filtre de qualité, filtre variants, et filtre sur les valeurs manquantes.

Caractérisation des données

La répartition des sites CpG observés ne correspond pas à la répartition dans le génome total, certaines régions étant significativement surreprésentées, grâce à la sélection effectuée par la technique RRBS. On note une surreprésentation des sites CpG dans les régions 5' UTR (+4 %), les exons (+6 %), et de façon plus prononcée, dans les promoteurs (+18 %), confirmant une concentration préférentielle des sites analysés dans des régions transcriptionnelles. En revanche, une sous-représentation significative et attendue a été observée dans les introns (-26 %) et également dans les régions 3' UTR (figure 10.A).

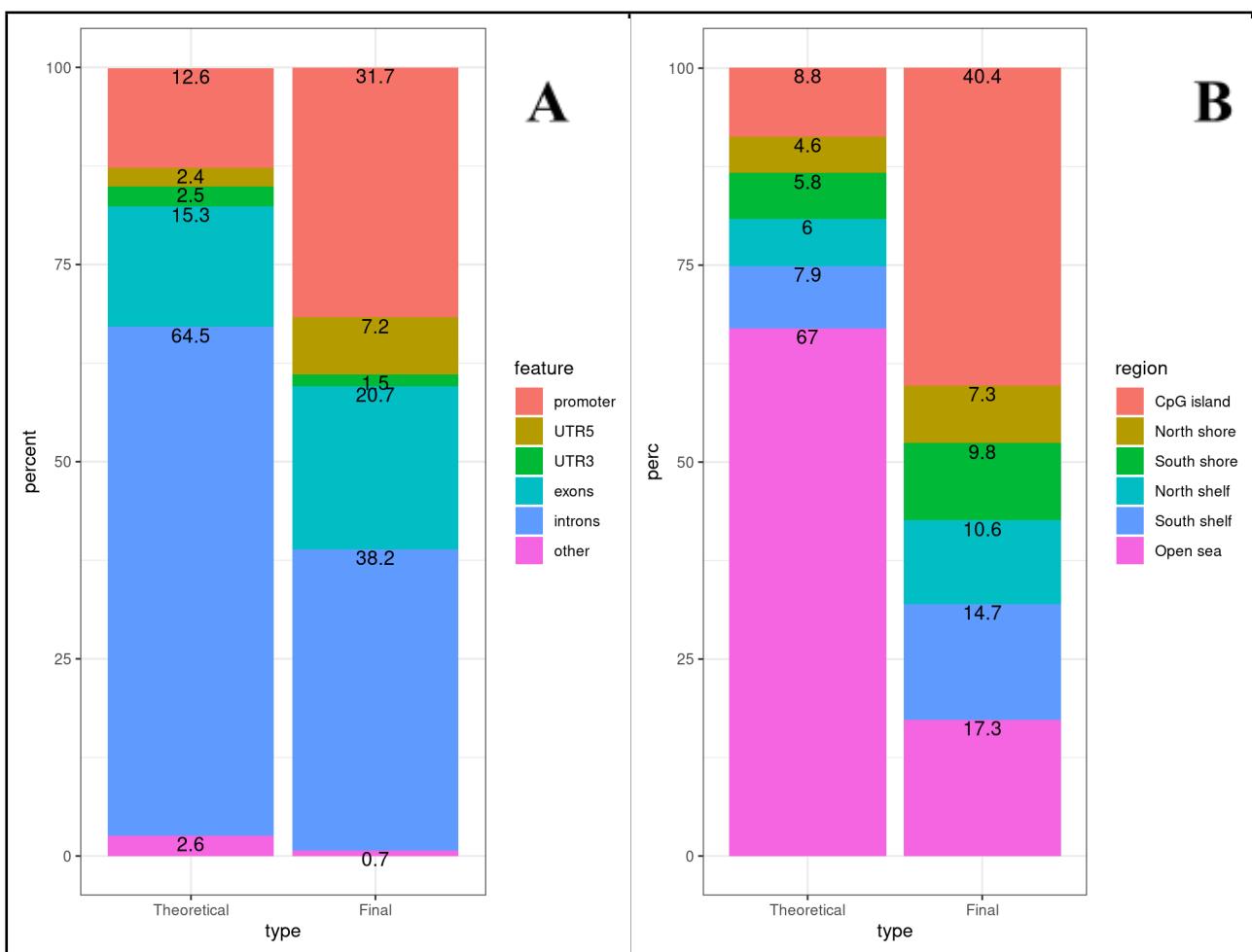


Figure 10 : Visualisation des proportions tout-génome (Theoretical) et observées dans notre jeu de données (Final) des sites CpG filtrés selon leur localisation sur le génome.

Le graphique A illustre la répartition des sites CpG selon leur position dans les différentes régions génomiques : intron, exon, 5' UTR, 3' UTR et autres.

Le graphique B montre la proportion des sites CpG en fonction de leur distance par rapport aux îlots CpG : îlot CpG, rive nord, rive sud, plateau nord, plateau sud, et extérieur de l'îlot.

En ce qui concerne les positions par rapport aux îlots CpG, on a caractérisé 5 régions proches de ces îlots CpG. On retrouve le plateau nord situé à la suite de l'îlot CpG et de taille de 2 kb en 3' vers 5', puis la rive nord de 2 kb qui précède cette région. A contrario, le plateau sud est du côté 5' en 3'

l'îlot CpG et de même taille. On retrouve ensuite la rive sud de 2 kb. Au-delà de ces régions, on identifie l'extérieur de l'îlot, "open sea".

On note une hausse marquée de la représentation des sites CG situés dans les îlots (+32%), tout comme dans les plateaux du Sud. En revanche, les CpG localisés en dehors des îlots, ainsi que dans les plateaux Nord et les rives Sud, sont sous-représentés (figure 10.B). On obtient donc bien l'enrichissement en îlots CpG attendu après une analyse RRBS.

En explorant davantage la méthylation entre les échantillons, on peut se rendre compte rapidement de variations.

On constate que pour les deux facteurs, on retrouve davantage de variation de méthylation pour les sites CpG avec des moyennes de méthylation aux alentours de 0.5 et que les écarts-types sont plus importants, suggérant deux groupes avec des taux de méthylation différents. De plus, on retrouve plus de différences de méthylation pour la condition cage-sol avec 782 CpG qui se trouvent avec un delta de taux de méthylation supérieur à 0.05 (figure 11.A), se confirmant avec une analyse globale réalisée sur toutes les données de méthylation avec un test de Wilcoxon d'hypothèses:

- **Hypothèse nulle (H_0)** : Les groupes d'échantillons proviennent de **distributions identiques**
- **Hypothèse alternative (H_1)** : Une distribution tend à avoir **des valeurs plus grandes** (ou plus petites) que l'autre.

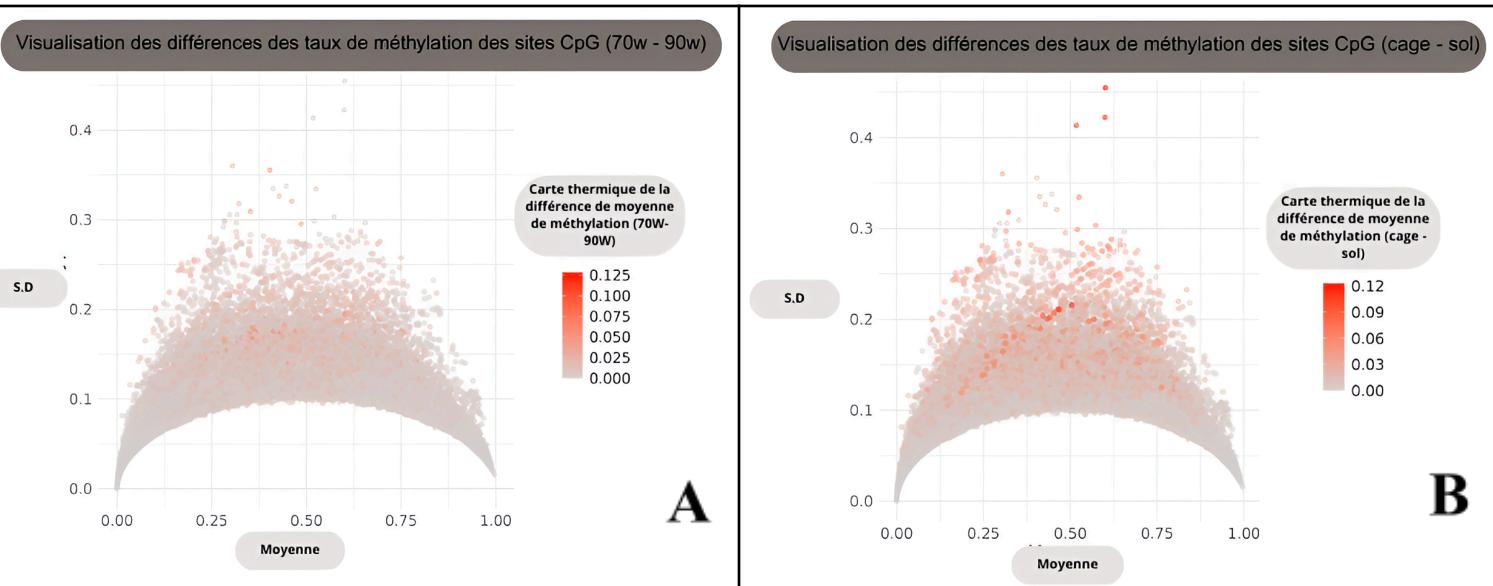


Figure 11: Visualisation de la différence des niveaux de méthylation des sites CpG en fonction de leur moyenne. Chaque point représente un site CpG, l'axe des abscisses correspondant au niveau moyen de méthylation, et l'axe des ordonnées représentant l'écart-type des valeurs de méthylation.

La coloration indique la différence de méthylation moyenne entre deux conditions :

- *A : Comparaison entre les âges W70 et W90 avec comme calcul : $|W70 - W90|$*
- *B : Comparaison entre les conditions d'élevage cage et sol avec comme calcul : $|cage - sol|$*

On retrouve des valeurs extrêmement significatives de p-value = 2.2e-16, traduisant une différence de distribution des taux de méthylation entre les conditions d'élevage. Tandis que pour la condition 70-90w, on retrouve seulement 196 CpG (figure 11.B) et une différence de globale de la distribution

de méthylation qui n'est pas significative entre ces deux groupes, démontrée par un test de Wilcoxon avec les mêmes hypothèses que précédemment de p-value = 0.57.

Analyse différentielle

Le modèle simple nous a permis d'identifier 10 680 DML différentes entre cage et sol pour les individus de 70 semaines et 48 210 pour les individus de 90 semaines. 6 401 DML sont communes à ces deux conditions.. On se rend compte d'une forte différence de DML et du nombre de ces DML entre le sous-échantillonnage à 70 semaines et 90 semaines.

Suite à ce nombre très important de DML, nous avons écarté des DML présentant une différence de variation de taux de méthylation faible mais aussi des DML qui ne sont pas robustes statistiquement avec une correction FDR de Benjamini-Hochberg. De ce fait, nous nous sommes concentrés sur les DML ayant une différence de taux de méthylation de plus de 5% entre nos deux conditions testées et ayant un FDR d'au moins 5%. On trouve 515 DML pour 70 semaines, 5 623 DML pour 90 semaines et 375 DML en commun.

On constate un déséquilibre du type de DML observées avec une sur-représentation des DML hyperméthylées pour la condition cage par rapport aux DML hypométhylées pour la condition cage avec l'application de ce modèle (Figure 13).

Dans le second modèle nous prenons en compte la complexité du design expérimental en incluant les interactions ainsi que l'ensemble des facteurs de l'étude. L'analyse de ce modèle révèle un nombre plus limité de DML pour l'interaction et le facteur d'âge. En revanche, le facteur élevage se distingue par un nombre élevé de DML et DMR hypométhylées ou hyperméthylées pour la condition cage avec 225 DML hypométhylées, 169 DML hyperméthylées, 9 DMR hypométhylées et 3 DMR hyperméthylées (figure 12), ce qui en fait un candidat pertinent pour l'analyse fonctionnelle et l'annotation biologique. On remarque aussi de manière plus générale une prépondérance de méthylation hypométhylées. Par ailleurs, une seule DML est commune entre les différents facteurs, elle se trouve être entre le facteur d'âge et d'interaction, cette DML est hypométhylé pour ces deux facteurs.

	élevage		âge		interaction	
DML	225	169	18	4	1	1
DMR	3	9	2	0	0	0

Figure 12: Tableau du nombre de DML et de DMR pour le modèle multifacteur avec l'effet élevage, âge et interaction avec en rouge les DML et DMR hyperméthylées pour la condition cage avec le facteur élevage et 90w pour le facteur âge, en bleu les DMR et DML hypométhylées pour ces conditions.

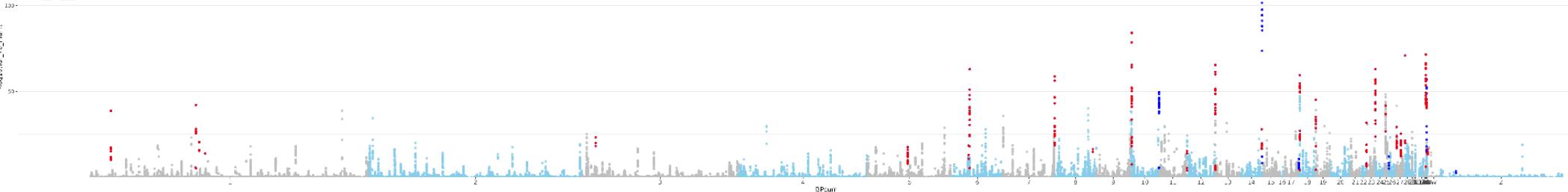
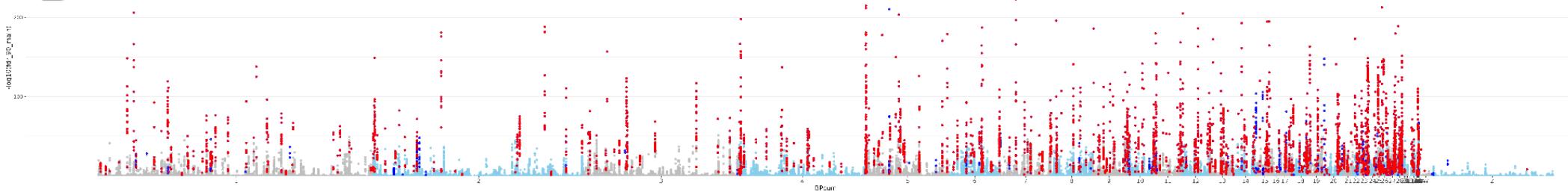
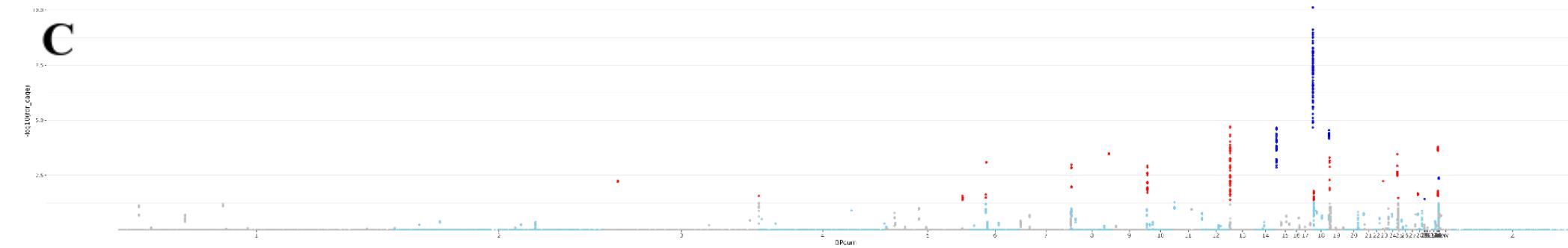
A**B****C**

Figure 13: Manhattan plot des DML obtenu à partir de DSS sur le sous-échantillonnage des échantillons présents à la semaine 70 pour le graphique A ou à la semaine 90 pour le graphique B, le graphique C est obtenu avec l'analyse de l'effet cage du modèle multifacteur. L'axe des abscisses représente la position sur les différents chromosomes et l'axe des ordonnées représente le $-\log_{10}(p\text{-value})$. Chaque point correspond à 1 CpG analysé, et chaque point en rouge ou en bleu marine est filtré avec un différentiel de méthylation de minimum 5% et un *fdr* de 5%. On retrouve des DML de couleur bleu marine qui correspondent aux DML hypométhylé tandis que les point rouge sont des DML hyperméthylés.

Analyse biologique

Une analyse fonctionnelle a été réalisée à partir des DML du modèle multifacteur avec le facteur cage. Sur l'ensemble des DML liés au facteur cage, 73 se trouvent localisés dans des régions géniques, tandis que 74 sont situés dans des promoteurs. L'analyse dans les régions géniques n'a révélé aucune catégorie significativement enrichie. Tandis que dans les promoteurs, on retrouve un enrichissement qui a mis en évidence deux gènes significativement enrichis: **VPS26A** (Vacuolar Protein Sorting-Associated Protein 26A) **qui est hyperméthylé** (adj.p-value = 0,0396) associé à des fonctions de transport dans l'endosome vers l'appareil de Golgi, et **SRA1** (Steroid Receptor RNA Activator 1) **qui est hyperméthylé** (adj.p-value = 0,0299) est un récepteur stéroïdien qui a une implication dans la régulation négative de la différenciation des cellules souches en myoblastes (figure 14).

Pour le facteur d'âge, on ne retrouve pas suffisamment de DML pour effectuer une analyse SEA, mais en regardant la co-localisation de ces DML avec les régions des gènes ou des promoteurs on retrouve 18 DML hypométhylées et présents dans 4 différents gènes: **NPAS3** (Neuronal PAS Domain Protein 3), **SPARC** (Secreted Protein Acidic And Cysteine Rich), **PPARA** (Peroxisome Proliferator Activated Receptor Alpha), **ALS2** (Alsin Rho Guanine Nucleotide Exchange Factor) et aucun présent dans des promoteurs. Avec l'interaction des deux facteurs, on trouve 1 DML présente dans le promoteur du gène hypométhylé **CORO1B** (Coronin 1B) et deux DML présentes dans la région génomique dont un gène qui est hypométhylé **ALS2** et un autre qui est hyperméthylé **CABP4** (Calcium Binding Protein 4). **ALS2** contient une DML présente à la fois dans les DML d'interaction et de facteur d'âge. Une liste exhaustive des DML, DMR et de leur position par rapport à des gènes et promoteurs est reportée dans le tableau en annexe.

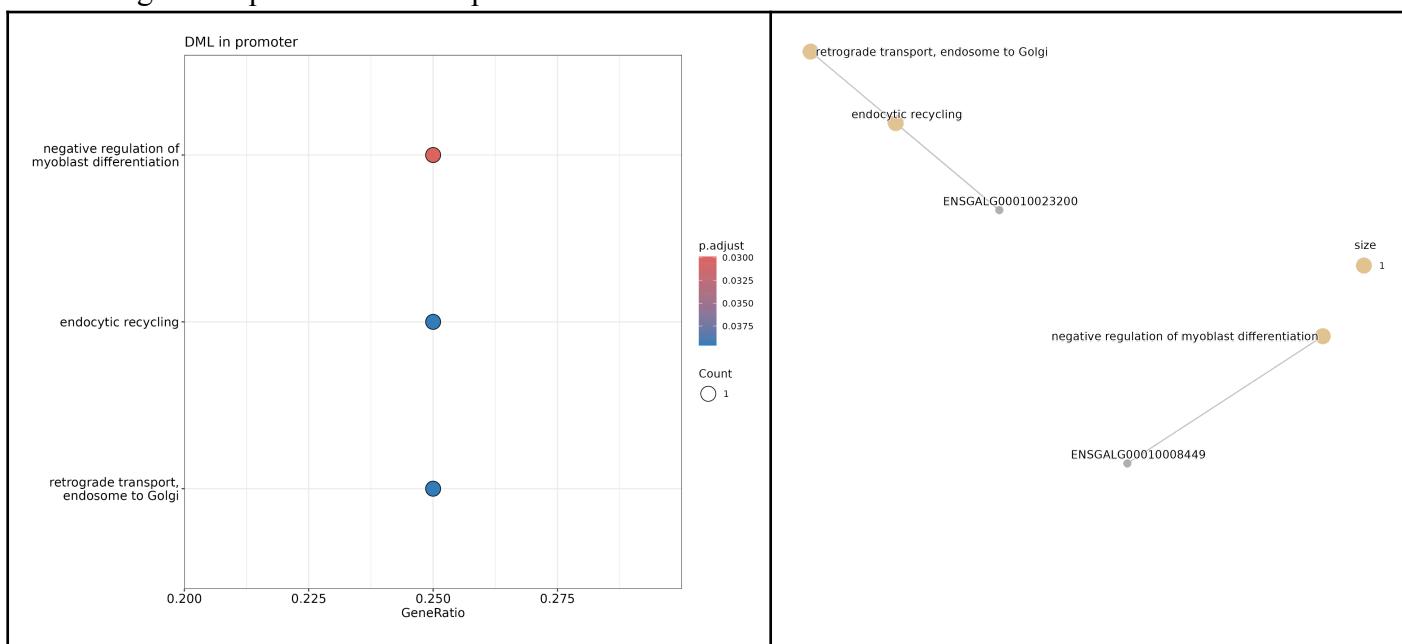


Figure 14 : Visualisation de l'enrichissement fonctionnel des DML associés au facteur élevage dans le modèle multifactoriel pour les promoteurs. À droite, un dotplot représentant la significativité des fonctions enrichies, visualisée sous forme de heatmap. À gauche, une représentation en réseau illustrant les liens entre les gènes impliqués et les fonctions enrichies.

Discussion

L'analyse de méthylation de l'ADN afin de caractériser l'âge ou le mode d'élevage des poules pondeuses a été réalisée sur du sang. L'extraction de cellules sanguines a été retenue du fait de la facilité du procédé et de la possibilité de la répétabilité de la prise de sang plusieurs fois dans la vie d'une poule. Ce type d'échantillonnage permet une analyse directe et sans biais de composition cellulaire du tissu analysé : la plupart des tissus sont hétérogènes, en comportant différents types cellulaires. Chez les oiseaux les globules rouges sont nucléés et constituent la vaste majorité des cellules sanguines, donc après une simple centrifugation nous pouvons disposer d'un tissu composé d'un type cellulaire unique. C'est particulièrement important pour les analyses de méthylation, le méthylome étant variable en fonction du type cellulaire. Par ailleurs, ces cellules illustrent l'effet des conditions environnementales: une adaptation au stress chronique des systèmes d'élevage en cage a été montrée comme induisant des différences de méthylation sur certaines régions génomiques qui reflètent une réponse aux stress environnementaux et d'isolement, ainsi qu'une diminution de la réponse du système immunitaire (*Pértille et al., 2017*).

A l'issue du pipeline d'analyse, nous obtenons un nombre total de 592 391 sites CpG uniques ce qui est satisfaisant, avec une profondeur de lecture moyenne correcte de 22,78 X, suite à l'application des différents filtres de qualité. De plus, les données révèlent une surreprésentation attendue des sites CpG dans les régions promotrices, ainsi qu'un enrichissement au sein des îlots CpG, ce qui est attendu et confirme la pertinence de notre analyse RRBS. Avec la réalisation de deux différents modèles, on constate que le modèle simple n'est pas un modèle qui prend suffisamment d'informations mises à notre disposition. Notamment, l'absence d'interaction entre nos facteurs se traduit par un nombre de DML très important pour le sous-échantillon qui comprend les échantillons de 90 semaines. Nous avons donc choisi de ne retenir que le modèle le plus élaboré, qui nous a permis de mettre en évidence 394 DML entre modes d'élevage et 22 entre les deux âges. A partir de la distance de la région annotée la plus proche de chaque DML, nous avons observé qu'une partie des DML qui se situent dans des régions intergéniques, se trouve à proximité de promoteurs ou de gènes $\pm 2000\text{pb}$, ces DML pourraient être intéressants à explorer, dans une analyse future, pour leur potentielle implication dans la régulation de certains gènes (tableau en annexe).

Il est important de rappeler que notre approche repose sur une analyse différentielle de la méthylation de l'ADN, considérée comme un régulateur de l'expression génique. Ainsi, les variations de méthylation observées peuvent refléter des altérations potentielles de l'activité fonctionnelle des gènes.

Nous avons effectué une analyse d'enrichissement pour prioriser les gènes les plus intéressants seulement pour le facteur d'élevage car ce facteur présente suffisamment de DML. On retrouve ainsi, pour l'impact du facteur de condition d'élevage, une hyperméthylation dans le sens de la condition cage pour le promoteur de **SRA1** et **VPS26A**, qui pourrait entraîner une réduction de l'expression des gènes associés et par conséquent altérer les fonctions biologiques associées. Chez la poule, **VPS26A** joue un rôle de transporteur vésiculaire ainsi que dans la formation des vésicules extracellulaires dans l'oviducte, une structure importante pour la formation de la coquille d'œuf (*Stapane et al., 2020*). Une répression de l'expression de ce gène pourrait entraîner une

fragilité dans la minéralisation des coquilles d'œuf. Une autre fonction de ce gène est un rôle dans le transport de récepteur immunoglobuline (*Meitern et al., 2014*), pouvant se traduire par une altération du système immunitaire de la poule.

De son côté, **SRA1** est plus complexe et code à la fois pour des formes codantes et non codantes. Il est notamment impliqué dans la différenciation négative des myoblastes, fonction qui ne semble pas en lien direct avec la modification du mode d'élevage entre les poules en cage et au sol (à partir de 17 semaines), alors que cette différenciation se termine à 21 jours. Une isoforme non codante de ce gène, qui est peu étudiée chez la poule, a été identifiée chez la souris et sa sous-expression pourrait causer des problèmes dans la métabolisation des lipides (*Muret et al., 2019*). Elle jouerait aussi un rôle dans le maintien de la fonction cardiaque et, par conséquent, une sous-expression de ce gène signifierait des dysfonctionnements contractiles du myocarde (*Friedrichs et al., 2009*).

Pour l'impact du facteur de l'âge on retrouve seulement des gènes hypométhylés pour la condition 90 semaines, l'interprétation est à prendre avec plus de recul, car nous n'avons pas pu réaliser d'analyse d'enrichissement du fait du nombre trop peu important de DML retrouvées.

On retrouve cependant des DML dans le gène **SPARC** qui est impliqué dans le développement de la rétine (*Kim et al., 1997*), mais notre dispositif ne devrait pas nous permettre de voir des gènes impliqués dans le développement précoce car la fin de développement de la rétine se termine au bout de 20–21 jours (*Zareen et al., 2011*), mais ce gène est aussi impliqué dans la calcification des os chez l'humain, pouvant alors induire une insuffisance en calcium (*Storoni et al., 2023*). **PPARA** est lui associé au métabolisme lipidique et une surexpression de ce gène entraînerait de l'obésité avec l'âge (*Suzuki et al., 2019*). Deux gènes du système nerveux ont été mis en évidence, **NPAS3** (*Shin and Kim, 2013*) qui est impliqué dans le développement des neurones et **ALS2** (*Hadano et al., 2007*) qui possède tout comme **NPAS3** une fonction de neurodégénérescence. **NPAS3**, en interaction avec un lncRNA, est impliqué dans la neurodégénérescence liée à l'âge (*Schröder et al., 2025*). **ALS2** a aussi été identifié dans l'analyse de l'interaction âge/élevage. Ce gène a été montré comme différentiellement méthylé en fonction de traumatismes subis ou non dans l'enfance chez l'homme, il pourrait constituer dans notre modèle une “mémoire” de la vie en cage ou au sol (*Labonté et al., 2012*). Enfin, pour l'impact de l'interaction entre nos deux facteurs, nous avons identifié une DML en overlap entre le promoteur du gène **CORO1B** et le gène **ALS2**. **CORO1B** possède une fonction dans le cytosquelette. Cette région présente un intérêt particulier : le DML présent ici se colocalise avec le promoteur du gène **CORO1B** sur le brin positif et le gène **ALS2** sur le brin négatif. La méthylation différentielle pourrait réguler de manière non spécifique (effet épigénétique potentiellement non ciblé) le gène **CORO1B** qui présente des fonctions moins cohérentes dans notre contexte que le gène **ALS2**, retrouvé différentiellement méthylé dans l'analyse du facteur d'âge. Toujours dans l'interaction entre les facteurs d'élevage et d'âge, on retrouve le gène **CaBP4** qui entraîne une diminution des liaisons synaptiques des photorécepteurs Ca²⁺ dans la rétine et dans la cochlée (*Haeseleer et al., 2013*), pouvant s'expliquer par une exposition répétée dans le temps à un éclairage qui ne présente pas de contraste et par conséquent à une réduction de la sensibilité à la lumière pour la poule.

Perspectives:

Dans le cadre de notre étude nous avons utilisé BISCUIT pour le calling des méthylations et l'estimation de leur taux, et DSS pour l'analyse différentielle. Nous savons qu'il existe d'autres outils pour réaliser cette analyse. En alternative à BISCUIT nous aurions pu utiliser nf-core/methylseq, qui se base sur Bismark (*Krueger and Andrews, 2011*) ou BWA-meth (*Kerns and Weber, 2025*) pour identifier les CpG méthylés et estimer leur taux de méthylation. Notre préférence s'est portée vers BISCUIT, un outil permettant à la fois l'estimation du taux de méthylation, mais aussi des variants génétiques, importants pour d'autres aspects du projet de recherche. Pour réaliser l'analyse différentielle, nous aurions pu choisir d'autres méthodes que DSS, telles que edgeR (*Robinson et al., 2010*) ou BSseq (*Hansen et al., 2012*), cependant DSS à l'avantage d'utiliser, dans son modèle, une distribution appropriée à nos données de méthylation mais aussi l'attribution de poids sur le taux de méthylation en fonction de la profondeur, c'est pourquoi notre choix s'est porté sur celui ci. Dans le cadre du projet GEroNIMO un benchmark de ces outils est prévu.

Dans la continuité de ce travail, une piste d'amélioration consisterait à explorer des modèles multivariés avec une variable concaténée de nos facteurs d'élevage et de l'âge, ce qui pourrait nous permettre de contourner les problèmes techniques liés à l'appariement partiel de nos données. Ce modèle pourrait nous permettre de prendre en compte l'intégralité des échantillons et donc d'utiliser toutes les informations mises à notre disposition.

Par ailleurs, l'utilisation d'une approche de type GSEA pourrait être une alternative intéressante pour notre enrichissement. Contrairement aux méthodes qui se basent sur des seuils stricts de p-value (SEA), le GSEA tient compte d'un classement de gènes selon une métrique a priori continue, lui conférant une meilleure puissance statistique pour détecter des petits enrichissements, notamment dans notre contexte où l'on obtient peu de DML.

En conclusion, de notre analyse différentielle portant sur la méthylation, l'analyse nous a permis de mettre en évidence l'influence combinée de l'âge et des conditions d'élevage sur la régulation de gènes liés à des fonctions biologiques essentielles, telles que le métabolisme, l'immunité, la minéralisation, le système nerveux et rétinien. Ces résultats mettent en perspective le rôle de la méthylation qui s'inscrit dans des processus d'adaptation des individus face à des stress liés à leur condition d'élevage et ou à l'âge, ainsi que de montrer des marqueurs moléculaires des potentielles vulnérabilités face aux environnements d'élevage, avec des implications directes sur le bien-être et les performances physiologiques des animaux.

Bibliographie :

A.

1. Al Adhami, H., Bardet, A.F., Dumas, M., Cleroux, E., Guibert, S., Fauque, P., Acloque, H., Weber, M., 2022. A comparative methylome analysis reveals conservation and divergence of DNA methylation patterns and functions in vertebrates. *BMC Biology* 20, 70. <https://doi.org/10.1186/s12915-022-01270-x>
2. Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
3. Attwood, J.T., Yung, R.L., Richardson, B.C., 2002. DNA methylation and the regulation of gene transcription. *CMLS, Cell. Mol. Life Sci.* 59, 241–257. <https://doi.org/10.1007/s00018-002-8420-z>

B.

4. Berger, Q., Bedere, N., Lagarrigue, S., Burlot, T., Le-Roy, P., Tribout, T., Zerjal, T., 2025. Unravelling the genetic architecture of persistence in production, quality, and efficiency traits in laying hens at late production stages. <https://doi.org/10.1101/2025.02.26.640268>
5. Bollati, V., Schwartz, J., Wright, R., Litonjua, A., Tarantini, L., Suh, H., Sparrow, D., Vokonas, P., Baccarelli, A., 2009. Decline in genomic DNA methylation through aging in a cohort of elderly subjects. *Mechanisms of Ageing and Development* 130, 234–239. <https://doi.org/10.1016/j.mad.2008.12.003>

C.

6. Commission to propose phasing out of cages for farm animals, 2021. European Commission. URL https://ec.europa.eu/commission/presscorner/detail/en/ip_21_3297 (accessed 5.25.25).

D.

7. Danecek P. et al. 2021. Twelve years of SAMtools and BCFtools. *Gigascience* 10, giab008. <https://doi.org/10.1093/gigascience/giab008>
8. Degalez, F., Bardou, P., Lagarrigue, S., 2024. GEGA (Gallus Enriched Gene Annotation): an online tool providing genomics and functional information across 47 tissues for a chicken gene-enriched atlas gathering Ensembl and Refseq genome annotations. *NAR Genomics and Bioinformatics* 6, lqae101. <https://doi.org/10.1093/nargab/lqae101>
9. Degalez, F et al. 2024. Enriched atlas of lncRNA and protein-coding genes for the GRCg7b chicken assembly and its functional annotation across 47 tissues. *Sci Rep* 14, 6588. <https://doi.org/10.1038/s41598-024-56705-y>

E.

10. Ewels, P., Magnusson, M., Lundin, S., Käller, M., 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32, 3047–3048. <https://doi.org/10.1093/bioinformatics/btw354>

F.

11. Feil, R., 2006. Environmental and nutritional effects on the epigenetic regulation of genes. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 600, 46–57.
<https://doi.org/10.1016/j.mrfmmm.2006.05.029>
12. Feng, H., Conneely, K.N., Wu, H., 2014. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Res* 42, e69.
<https://doi.org/10.1093/nar/gku154>
13. Friedrichs, F et al. 2009. HBEGF, SRA1, and IK: Three cosegregating genes as determinants of cardiomyopathy. *Genome Res.* 19, 395–403. <https://doi.org/10.1101/gr.076653.108>

G.

14. Gaillac, R., Marbach, S., 2021. The carbon footprint of meat and dairy proteins: A practical perspective to guide low carbon footprint dietary choices. *Journal of Cleaner Production* 321, 128766.
<https://doi.org/10.1016/j.jclepro.2021.128766>
15. Gryzinska, M., Blaszcak, E., Strachecka, A., Jezewska-Witkowska, G., 2013. Analysis of Age-Related Global DNA Methylation in Chicken. *Biochem Genet* 51, 554–563. <https://doi.org/10.1007/s10528-013-9586-9>

H.

16. Hadano, S., Kunita, R., Otomo, A., Suzuki-Utsunomiya, K., Ikeda, J.-E., 2007. Molecular and cellular function of ALS2/alsin: Implication of membrane dynamics in neuronal development and degeneration. *Neurochemistry International*, The 50th Anniversary of the Japanese Society for Neurochemistry 51, 74–84.
<https://doi.org/10.1016/j.neuint.2007.04.010>
17. Haeseleer, F., Sokal, I., Gregory, F.D., Lee, A., 2013. Protein Phosphatase 2A Dephosphorylates CaBP4 and Regulates CaBP4 Function. *Investigative Ophthalmology & Visual Science* 54, 1214–1226.
<https://doi.org/10.1167/iovs.12-11319>
18. Hansen, K.D., Langmead, B., Irizarry, R.A., 2012. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology* 13, R83. <https://doi.org/10.1186/gb-2012-13-10-r83>
19. Hao Wu, 2025. The DSS User's Guide [WWW Document] URL
<https://bioconductor.riken.jp/packages/3.10/bioc/vignettes/DSS/inst/doc/DSS.html>

K.

20. Kerns, E.V., Weber, J.N., 2025. Variable performance of widely used bisulfite sequencing methods and read mapping software for DNA methylation. <https://doi.org/10.1101/2025.03.14.643302>
21. Kim, S.Y., Ondhia, N., Vidgen, D., Ringuette, M., Malaval, L., Kalnins, V.I., 1997. Spatiotemporal distribution of SPARC/Osteonectin in Developing and Mature Chicken Retina. *Experimental Eye Research* 65, 681–689.
<https://doi.org/10.1006/exer.1997.0377>
22. Krueger, F., Andrews, S.R., 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 27, 1571–1572. <https://doi.org/10.1093/bioinformatics/btr167>

L.

23. Labonté, B., Suderman, M., Maussion, G., Navaro, L., Yerko, V., Mahar, I., Bureau, A., Mechawar, N., Szyf, M., Meaney, M.J., Turecki, G., 2012. Genome-wide Epigenetic Regulation by Early-Life Trauma. *Arch Gen Psychiatry* 69, 722–731. <https://doi.org/10.1001/archgenpsychiatry.2011.2287>

24. Li, Q., Li, N., Hu, X., Li, J., Du, Z., Chen, L., Yin, G., Duan, J., Zhang, H., Zhao, Y., Wang, J., 2011. Genome-Wide Mapping of DNA Methylation in Chicken. PLOS ONE 6, e19428. <https://doi.org/10.1371/journal.pone.0019428>

M.

25. Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, 10–12. <https://doi.org/10.14806/ej.17.1.200>
26. Meitern, R., Andeson, R., Höök, P., 2014. Profile of whole blood gene expression following immune stimulation in a wild passerine. BMC Genomics 15, 533. <https://doi.org/10.1186/1471-2164-15-533>
27. Muret, K., Désert, C., Lagoutte, L., Boutin, M., Gondret, F., Zerjal, T., Lagarrigue, S., 2019. Long noncoding RNAs in lipid metabolism: literature review and conservation analysis across species. BMC Genomics 20, 882. <https://doi.org/10.1186/s12864-019-6093-3>

N.

28. Nakabayashi, K., Yamamura, M., Hasegawa, K., Hata, K., 2023. Reduced Representation Bisulfite Sequencing (RRBS). Methods Mol Biol 2577, 39–51. https://doi.org/10.1007/978-1-0716-2724-2_3

P.

29. Pétille, F., Brantsæter, M., Nordgreen, J., Coutinho, L.L., Janczak, A.M., Jensen, P., Guerrero-Bosagna, C., 2017. DNA methylation profiles in red blood cells of adult hens correlate with their rearing conditions. Journal of Experimental Biology 220, 3579–3587. <https://doi.org/10.1242/jeb.157891>

R.

30. Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140. <https://doi.org/10.1093/bioinformatics/btp616>
31. Romero, I.G., Ruvinsky, I., Gilad, Y., 2012. Comparative studies of gene expression and the evolution of gene regulation. Nat Rev Genet 13, 505–516. <https://doi.org/10.1038/nrg3229>
32. Roussot, O., Feve, K., Plisson-Petit, F., Pitel, F., Faure, J.-M., Beaumont, C., Vignal, A., 2003. AFLP linkage map of the Japanese quail *Coturnix japonica*. Genetics Selection Evolution 35, 559. <https://doi.org/10.1186/1297-9686-35-6-559>

S.

33. Schröder, S., Sakib, M.S., Krüger, D.M., Pena, T., Burkhardt, S., Schütz, A.-L., Sananbenesi, F., Fischer, A., 2025. LncRNA 3222401L13Rik Is Upregulated in Aging Astrocytes and Regulates Neuronal Support Function Through Interaction with Npas3. Noncoding RNA 11, 2. <https://doi.org/10.3390/ncrna11010002>
34. Shin, J., Kim, J., 2013. Novel alternative splice variants of chicken NPAS3 are expressed in the developing central nervous system. Gene 530, 222–228. <https://doi.org/10.1016/j.gene.2013.08.024>
35. Smith, J. et al. D.W., 2000. Differences in gene density on chicken macrochromosomes and microchromosomes. Animal Genetics 31. <https://doi.org/10.1046/j.1365-2052.2000.00565.x>

36. Smith, T., Heger, A., Sudbery, I., 2017. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* 27, 491–499.
<https://doi.org/10.1101/gr.209601.116>
37. Stapane, L., Roy, N.L., Ezagal, J., Rodriguez-Navarro, A.B., Labas, V., Combes-Soia, L., Hincke, M.T., Gautron, J., 2020. A new model for vertebrate mineralization via stabilized amorphous calcium carbonate for avian eggshell formation. <https://doi.org/10.1101/2020.04.08.031989>
38. Storoni, S. et al. 2023. Novel pathogenic variants in SPARC as cause of osteogenesis imperfecta: Two case reports. *Eur J Med Genet* 66, 104857. <https://doi.org/10.1016/j.ejmg.2023.104857>
39. Suzuki, S., Kobayashi, M., Murai, A., Tsudzuki, M., Ishikawa, A., 2019. Characterization of Growth, Fat Deposition, and Lipid Metabolism-Related Gene Expression in Lean and Obese Meat-Type Chickens. *The Journal of Poultry Science* 56, 101–111. <https://doi.org/10.2141/jpsa.0180064>

T.

40. Tulchinsky, T.H., 2010. Micronutrient Deficiency Conditions: Global Health Issues. *Public Health Rev* 32, 243–255. <https://doi.org/10.1007/BF03391600>

W.

41. World Population Prospects: The 2017 Revision Volume II: Demographic Profiles | Population Division [WWW Document], URL
<https://www.un.org/development/desa/pd/content/world-population-prospects-2017-revision-volume-ii-demographic-profiles>.
42. Wu, H., Wang, C., Wu, Z., 2013. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics* 14, 232–243. <https://doi.org/10.1093/biostatistics/kxs033>

Y.

43. Yu, G., Wang, L.-G., Han, Y., He, Q.-Y., 2012. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology* 16, 284–287.
<https://doi.org/10.1089/omi.2011.0118>

Z.

44. Zareen, N., Khan ,M. Y., and Minhas, L.A., 2011. Histological stages of retinal morphogenesis in chicken – a descriptive laboratory research. *Italian Journal of Zoology* 78, 45–52.
<https://doi.org/10.1080/11250003.2010.487075>
45. Zhao, Y. (Ed.), 2022. Housing Environment and Farm Animals' Well-Being. MDPI - Multidisciplinary Digital Publishing Institute. <https://doi.org/10.3390/books978-3-0365-4585-1>
46. Zhou, W et al. 2024. BISCUIT: an efficient, standards-compliant tool suite for simultaneous genetic and epigenetic inference in bulk and single-cell studies. *Nucleic Acids Res* 52, e32.
<https://doi.org/10.1093/nar/gkae097>

Annexe

Introduction

Une des réalisations créatives majeures de cette étude a été le benchmarking d'un seuil de profondeur maximale de lecture, spécifiquement conçu pour s'adapter à notre physiologie particulière de distribution des données de profondeur issues du séquençage. En effet, des variations locales de profondeur au sein des chromosomes sont présentes du fait de la présence de CpG sur des séquences qui se répètent dans le génome.

Seuil de filtration de profondeur maximale (Matériel et Méthode)

Cette variabilité rend nécessaire une approche de filtrage personnalisée afin d'éviter de conserver une profondeur trop importante pour des CpG qui se localisent sur des séquences répétées dans le génome. Pour répondre à cette problématique, trois stratégies de filtration ont été développées et testées sur un sous-échantillon représentatif du jeu de données global. Chaque filtre visait à exclure les sites anormalement couverts tout en préservant une distribution équilibrée des sites CpG analysés.

On retrouve alors ces trois seuils :

- Un seuil uniforme, qui se base sur le quantile de profondeur globale.
- Un seuil par chromosome, qui se base sur le quantile de profondeur par chromosome.
- Un seuil personnalisé, qui ajuste le quantile de profondeur par chromosome automatiquement. Ce seuil, plus complexe que les autres, repose sur plusieurs propriétés mathématiques liées aux dynamiques des courbes de distribution des profondeurs

Ce filtre est composé de deux étapes :

- La première étape repose sur l'analyse de la dérivée absolue de la courbe de densité des profondeurs par chromosome. L'idée est de suivre l'évolution de la densité de profondeur des lectures. Lorsque la densité varie fortement, la dérivée est élevée. En revanche, lorsque la densité devient stable ou diminue lentement (ce qui reflète une phase de raréfaction des profondeurs élevées), la dérivée devient très faible. Un seuil de 0,000001 a été fixé pour repérer ce point d'inflexion.

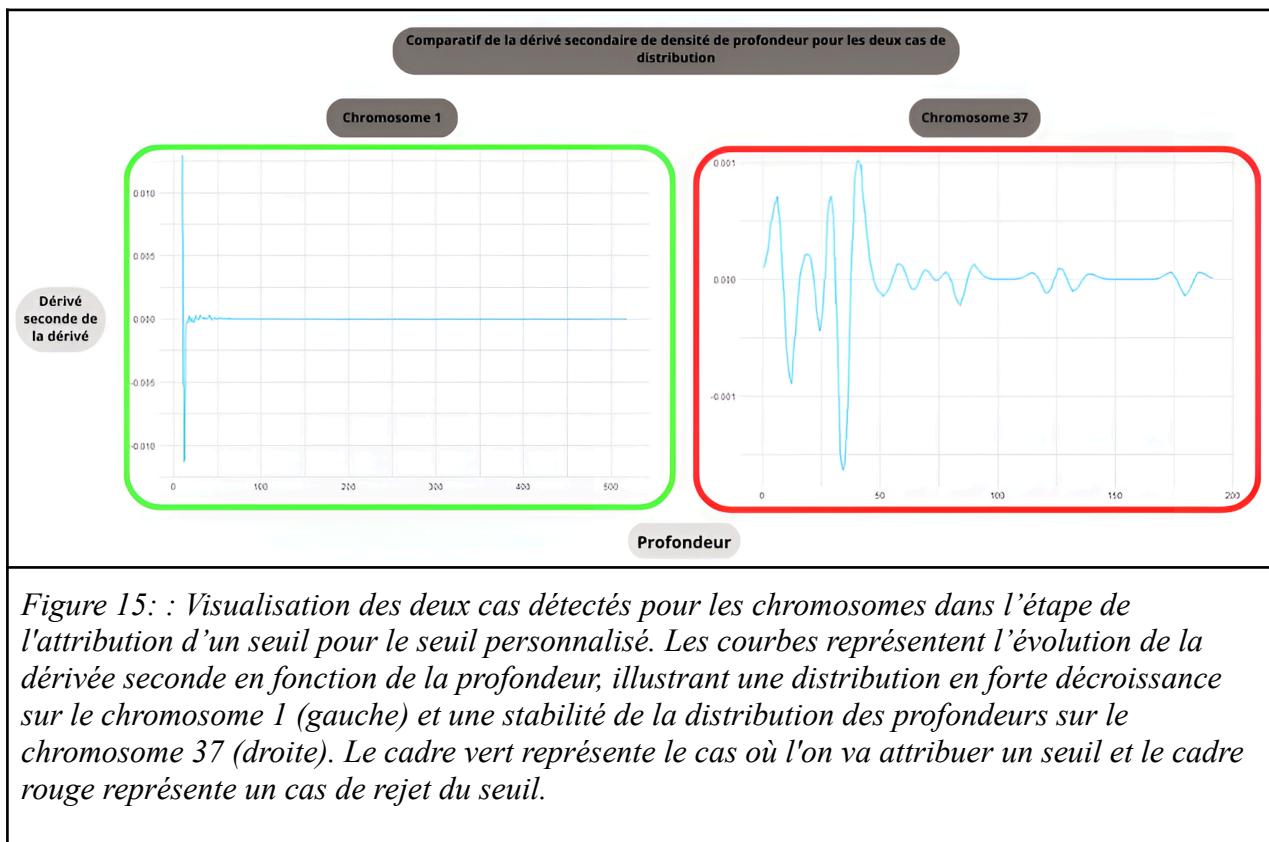
Cette méthode permet de réduire le bruit dû aux différences de profondeur entre échantillons et d'obtenir un seuil plus précis que si l'on utilisait directement la courbe de densité. En résumé, la dérivée absolue sert à détecter les petites variations dans la distribution des profondeurs, ce qui aide à définir un seuil objectif de filtrage.

- La seconde étape vise à supprimer l'attribution d'un seuil quand l'on observe une distribution de densité de profondeur répartie de façon relativement uniforme le long du chromosome ce qui traduit un séquençage peu important sur ce chromosome ou encore un chromosome faiblement riche en CpG. Pour détecter ce type de

distribution, nous utilisons la dérivée seconde de la courbe de densité, avec un seuil minimal fixé à -0,0018.

Cette dérivée permet d'évaluer la concavité de la courbe : une valeur fortement négative indique une chute marquée de la densité (donc une distribution non uniforme), tandis qu'une valeur proche de zéro indique une évolution plus stable, voire stationnaire, c'est ce cas que nous recherchons. De plus, nous nous concentrerons uniquement sur les zones à concavité négative, car en raison de la nature du séquençage, la distribution de la profondeur ne peut que décroître à mesure que la profondeur augmente.

En résumé, la dérivée seconde nous aide à repérer les profils où la densité décroît peu, ce qui traduit une répartition relativement uniforme des profondeurs. Cela permet d'éviter d'appliquer un seuil inadapté dans ces cas particuliers (voir Figure 15).



Benchmarking du seuil maximum (Résultats)

Dans le cadre du benchmarking des méthodes de détermination du seuil de profondeur maximale, trois méthodes ont été appliquées dont une complètement personnalisé pour cette

analyse. L'approche personnalisée a été comparée à une méthode basée sur le quantile par chromosome.

L'approche personnalisée, fondée sur l'analyse de la dérivée seconde de la densité de profondeur, a conduit à l'exclusion de cinq chromosomes (mitochondrie, 33, 32, 31, 29, 16 et 37) qui ont été jugés avoir une distribution trop uniforme : pour ces chromosomes, aucun seuil maximal n'a été appliqué. En regardant en détail le ratio CpG/taille du chromosome, on remarque en effet que les chromosomes qui ont été exclus du seuil de profondeur maximum comprennent un ratio anormalement faible par rapport au ratio qui augmente en fonction inverse de la taille des chromosomes (*Smith et al., 2000*) car on retrouve un enrichissement de gènes sur les macrochromosomes, donc par effet indirect une augmentation de CpG sur les macrochromosomes. Par exemple, on constate un ratio très peu élevé pour les chromosomes 33 et 37 qui présentent un ratio de 0.06% pour le chromosome 33 et 0.19% pour le chromosome 37 (figure 16).

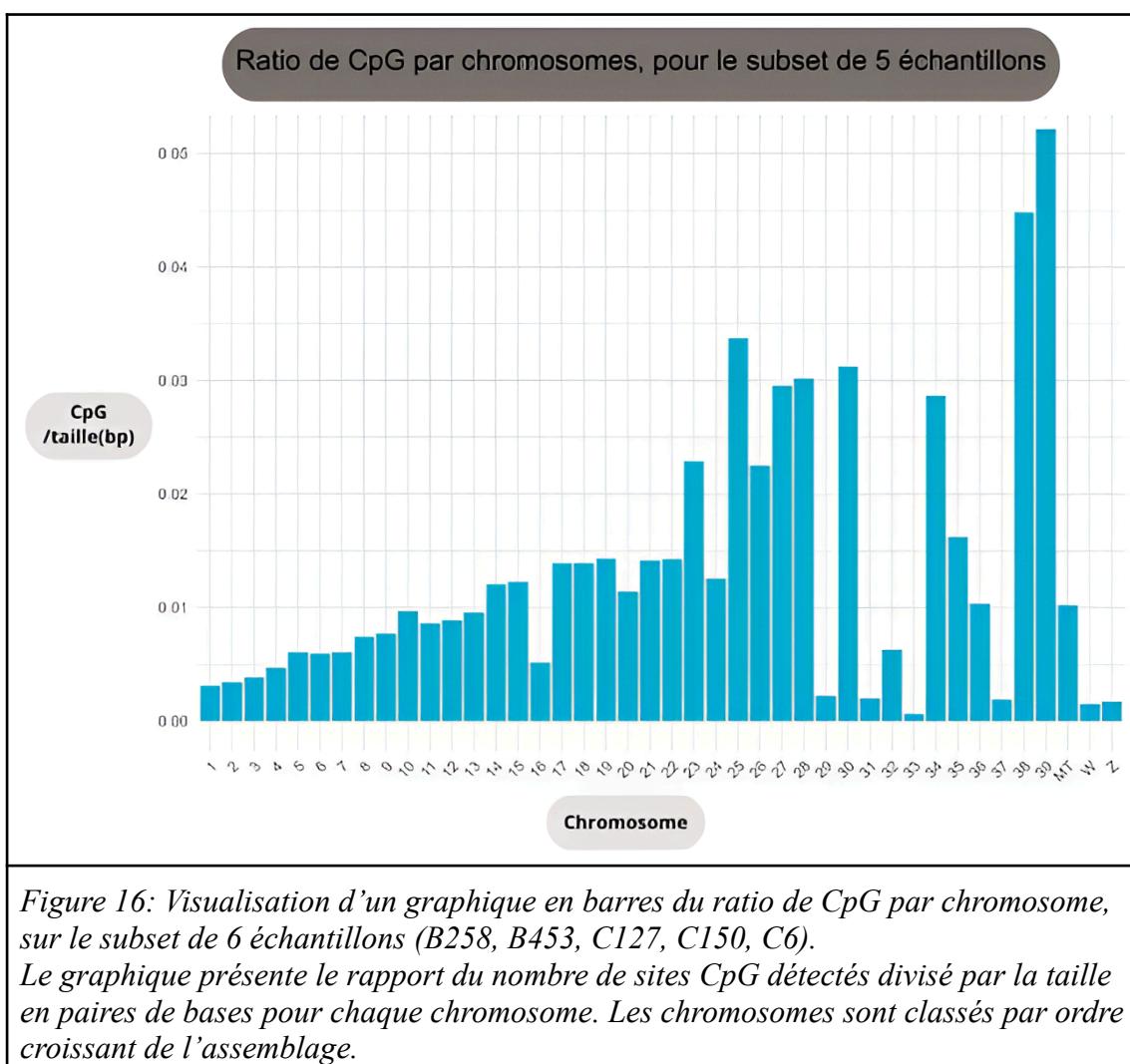


Figure 16: Visualisation d'un graphique en barres du ratio de CpG par chromosome, sur le subset de 6 échantillons (B258, B453, C127, C150, C6).

Le graphique présente le rapport du nombre de sites CpG détectés divisé par la taille en paires de bases pour chaque chromosome. Les chromosomes sont classés par ordre croissant de l'assemblage.

Maintenant, en comparant les deux méthodes de seuil, un échantillon de chromosomes a été sélectionné (qui représentent des cas particuliers de distributions de profondeur sur leur chromosomes, avec le chromosome 37, 7, 16 et une distribution attendue avec le chromosome 19 (figure 17), on note des disparités significatives entre les deux techniques. Par exemple, pour le chromosome 7, la méthode personnalisée a fixé une profondeur maximale de 98 lectures, tandis que

l'approche basée sur le quantile 1/1000 par chromosome était de 88. En revanche, une forte divergence a été notée sur le chromosome 16, avec un seuil de 270 dans la méthode personnalisée comparativement à 1850 pour le quantile. Concernant le chromosome 19, les deux approches ont produit des résultats comparables (100 contre 93), indiquant une concordance dans des conditions de répartition plus prévisibles.

Enfin, pour le chromosome 37, la définition d'un seuil a été exclue via la méthode personnalisée, alors que le seuil par quantile a été estimé à 160 (figure 17).

Conclusion du seuil maximum (Discussion)

Dans la plupart des cas, le filtre classique (quantile 1/1000 global) donnait des résultats similaires à notre filtre personnalisé sur les chromosomes qui ne présentent pas de particularité de distribution de profondeur (ce qui représente la majorité des chromosomes) mais avec tout de même moins de précision sur la définition du seuil maximum en faisant perdre plus de profondeur que nécessaire car on tend à garder un maximum d'informations de méthylation. Pour les cas où l'on retrouve un site CpG sur une séquence fortement répétée dans le génome (chromosome 16), la méthode qui repose sur le quantile ne prend pas correctement en compte ce facteur et néglige le seuil maximum contrairement à la méthode personnalisée. De plus, la méthode quantile se réalise sur tous les chromosomes même les chromosomes qui ne sont pas adaptés à un seuil maximum.

Nous avons quand même choisi d'utiliser un seuil de quantile maximum de 1/1000 de la profondeur pour la suite des analyses.

Le filtre quantile possède l'avantage d'être moins gourmand en temps, est automatique alors que le seuil personnalisé demande un suivi pour fixer une valeur seuil général en fonction de l'organisme et de la profondeur séquencée, en ressource computationnelle et présente aussi le bénéfice d'être une méthode déjà utilisée dans d'autres études.

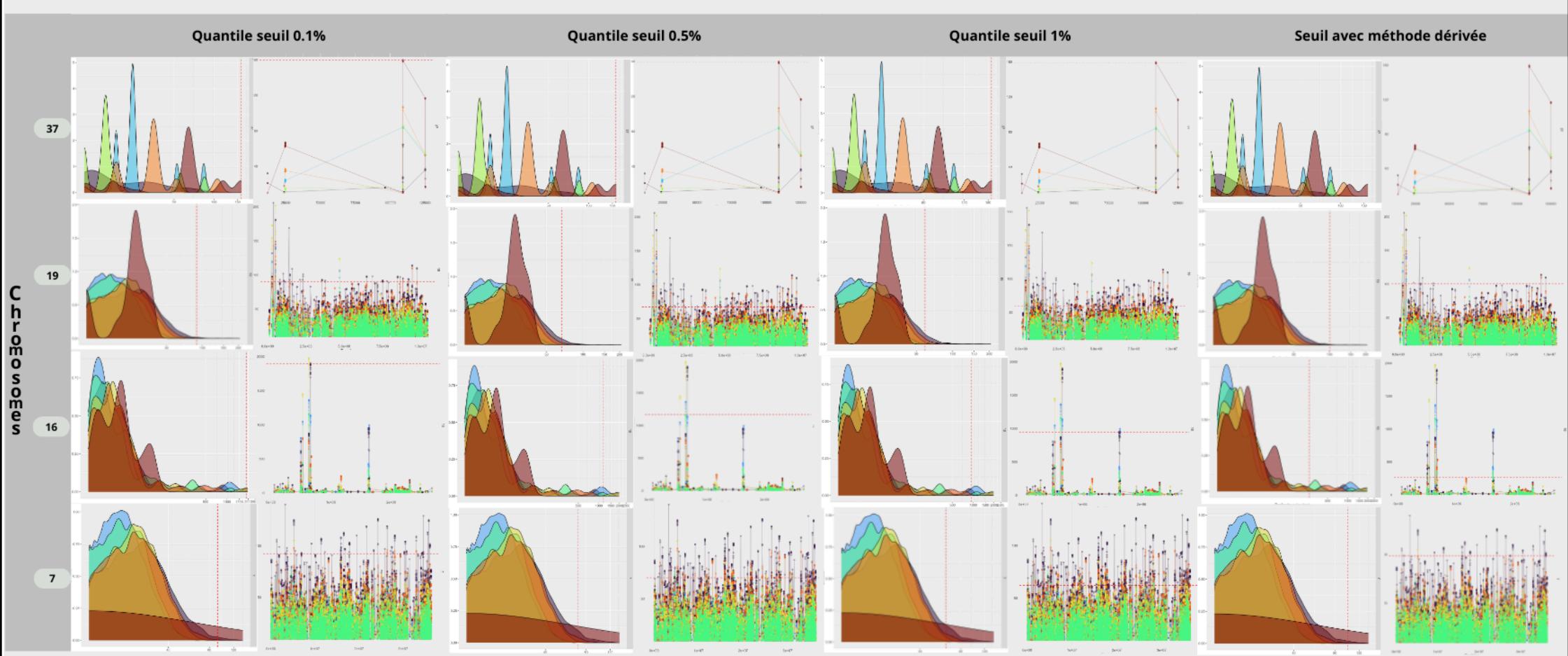


Figure 17 : Comparaison des méthodes de seuil de profondeur maximale pour le filtre de profondeur maximum, avec le seuil personnalisé et le seuil sur le quantile 1/1000 par chromosome sur les chromosomes 7, 16, 19 et 37. Chaque ligne correspond à un chromosome, et chaque colonne (des deux graphiques) à une méthode de sélection du seuil : quantiles (0,5 %, 1 %, 1,5 %) et méthode basée sur la dérivée (méthode de seuil personnalisée). Pour chaque combinaison, le premier graphique représente un plot de densité sur la profondeur maximum avec une ligne en pointillé rouge qui permet de visualiser le seuil appliqué. Le deuxième graphique représente la distribution de ces profondeurs sur le chromosome, avec aussi une ligne en pointillé qui représente le seuil appliqué. Chaque couleur représente un échantillon différent (B258, B453, C127, C150, C6) et chaque point un site CpG.

Barcodes

barcode_num	barcodeSequence	barcode_num	barcodeSequence	barcode_nu m	barcodeSequence
1	AACAGGTT+CTTGGTAT	33	CCATTCGA+CGGACAAC	65	GGCTTAAG+GGTCACGA
2	AACGTTCC+AGTACTCC	34	CCGCGGTT+CTAGCGCT	66	GGTACCTT+GACGTCTT
3	AACTGTAG+TGC GGCGT	35	CCGTGAAG+ATCCACTG	67	GGTCACGA+CATAATAC
4	AAGATACT+ATGTAAGT	36	CCTTCACC+GGAGCGTC	68	GGTGAACC+TCCAACGC
5	AAGTCCAA+TACTCATA	37	CGCTATGT+GTGTCGGA	69	GTATGTT+TTCCTGTT
6	AATCCGGA+AACTGTAG	38	CGGAACTG+TCGTAGTG	70	GTCGGAGC+TTATAACC
7	AATGCCTC+TGGATCGA	39	CGGACAAC+AATCCGGA	71	GTCTACAC+ACCTTGGC
8	ACACTAAG+ATATGGAT	40	CGGCGTGA+ACAGGCAC	72	GTGAATAT+GAATGAGA
9	ACAGGCCG+AGGCAGAG	41	CGTCTGCG+ATTGTGAA	73	GTGTCGGA+GCGCAAGC
10	ACCTTGGC+ATGAGGCC	42	CGTTAGAA+GACCTGAA	74	GTTCCAAT+GCAGAATT
11	ACGCACCT+CCTTCACC	43	CTACGACA+TTGGACTC	75	TAAGGTCA+CTACGACA
12	ACTAAGAT+CCGCGGTT	44	CTAGCGCT+GTCTACAC	76	TAAGTGGT+GGCTTAAG
13	ACTCGTGT+GTTCCAAT	45	CTCACCAA+TTGCCTAG	77	TAATACAG+GTGAATAT
14	AGCCTCAT+TCTCTACT	46	CTCTCGTC+AGGTTATA	78	TACCGAGG+AGTTCAAG
15	AGCTCGCT+GATTCTGC	47	CTGCTTCC+GATCTATC	79	TACTCATA+GCCACAGG
16	AGGCAGAG+GGCATTCT	48	CTTGGTAT+GGACTTGG	80	TATCGCAC+ACACTAAG
17	AGGTTATA+CGGAACGT	49	GAACCGCG+TAAGGTCA	81	TCATCCTT+AGCTCGCT
18	AGTACTCC+AACAGGTT	50	GAATGAGA+AATGCCTC	82	TCCAACGC+AAGTCCAA
19	AGTCAGG+TCTGTTGG	51	GACCTGAA+CTCACCAA	83	TCGATATC+ACTCGTGT
20	ATATCTCG+ACTAAGAT	52	GACGTCTT+GGTGAACC	84	TCGTAGTG+CCAAGTCT
21	ATATGGAT+TAATACAG	53	GATCTATC+AGCCTCAT	85	TCTCTACT+GAACCGCG
22	ATCCACTG+ACGCACCT	54	GATTCTGC+CTCTCGTC	86	TCTGTTGG+CCATTGCA
23	ATGAGGCC+CAATTAAC	55	GCAATGCA+AACGTTCC	87	TGCGAGAC+CAACAATG
24	ATGGCATG+GGTACCTT	56	GCACGGAC+TGCGAGAC	88	TGCGGCAG+TACCGAGG
25	ATGTAAGT+CATAGAGT	57	GCAGAATT+TGGCCGGT	89	TGGATCGA+TATCGCAC
26	ATTGTGAA+GCAATGCA	58	GCCACAGG+ATGGCATG	90	TGGCCGGT+GCGCTCTA
27	CAACAATG+CCGTGAAG	59	GCGCAAGC+CGGCGTGA	91	TGGTGGCA+TTACAGGA
28	CAAGCTAG+CGCTATGT	60	GCGCTCTA+GTCGGAGC	92	TTACAGGA+GCTTGTCA
29	CAATTAAC+ATATCTCG	61	GCTTGTCA+GTATGTT	93	TTATAACC+TCGATATC
30	CATAATAC+CGTTAGAA	62	GGACTTGG+CGTCTGCG	94	TTCCCTGTT+AAGATACT
31	CATAGAGT+TGGTGGCA	63	GGAGCGTC+GCACGGAC	95	TTGCCTAG+TAAGTGGT
32	CCAAGTCT+TCATCCTT	64	GGCATTCT+CAAGCTAG	96	TTGGACTC+CTGCTTCC

Abréviations

ADN	<i>Acide Désoxyribonucléique</i>	ORA	<i>Over-Representation Analysis</i>
ARN	<i>Acide Ribonucléique</i>	PCR	<i>Polymerase Chain Reaction</i>
CpG	<i>Cytosine-phosphate-Guanine</i>	RRBS	<i>Reduced Representation Bisulfite Sequencing</i>
DML	<i>Differentially Methylated Loci</i>	SEA	<i>Set Enrichment Analysis</i>
DMR	<i>Differentially Methylated Region</i>	UDI	<i>Unique Dual Index</i>
GSEA	<i>Gene Set Enrichment Analysis</i>	UMI	<i>Unique Molecular Identifier</i>
ORA	<i>Over-Representation Analysis</i>	UTR	<i>Untranslated Region</i>

Tableau de DML

Le tableau de DML est disponible en ligne à partir de ce lien et annoté avec la version 114 d'ensembl:

https://docs.google.com/spreadsheets/d/1b1AzEzQ4q3382lSV1jiO69ZwnGsH1_OEqC1JuWUoh0A/edit?usp=sharing