HAL open science

Identification de biomarqueurs de la qualité de la semence bovine, modélisation statistique et hypothèses biologiques.

Valentin Costes

To cite this version:

Valentin Costes. Identification de biomarqueurs de la qualité de la semence bovine, modélisation statistique et hypothèses biologiques.. Sciences du Vivant [q-bio]. Université Paris Saclay, 2022. Français. NNT : . tel-04534626

HAL Id: tel-04534626

https://hal.inrae.fr/tel-04534626

Submitted on 5 Apr 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Public Domain

Identification de biomarqueurs de la qualité de la semence bovine, modélisation statistique et hypothèses biologiques

Identification of biomarkers for bovine semen quality, statistical modelling and biological hypotheses

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 581 : Agriculture, Alimentation, Biologie, Environnement, Santé (ABIES)
Spécialité de doctorat : Biologie de la reproduction
Graduate School : Biosphera. Référent : AgroParisTech

Thèse préparée dans les unités de recherche **GABI** (Université Paris-Saclay, INRAE, AgroParisTech) et **BREED** (Université Paris-Saclay, UVSQ, INRAE) sous la direction de **Florence JAFFREZIC**, Directrice de recherche, le co-encadrement de **Hélène KIEFER,** Chargée de recherche et la co-supervision de **Laurent SCHIBLER**, Responsable Développement et Innovation (Entreprise Allice)

Thèse soutenue à Paris-Saclay, le 24 juin 2022, par

Valentin COSTES

Composition du Jury

Joël DREVET Professeur, Université Clermont Auvergne	Président
Sandrine LAGARRIGUE Professeure, Institut Agro Rennes-Angers	Rapporteur & Examinatrice
Julie COCQUET Chargée de recherche (HDR),INSERM (Sorbonne Paris-Cité)	Rapporteur & Examinatrice
David MAKOWSKI Directeur de recherche, INRAE (Université Paris-Saclay)	Examinateur
Florence JAFFREZIC Directrice de recherche, INRAE (Université Paris-Saclay)	Directrice de thèse

THESE DE DOCTORAT

NNT : 2022UPASB039

ÉCOLE DOCTORALE
Agriculture, alimentation, biologie, environnement, santé (ABIES)

université PARIS-SACLAY

Titre : Identification de biomarqueurs de la qualité de la semence bovine, modélisation statistique et hypothèses biologiques.

Mots clés : Epigénétique, Biomarqueurs, Intégration de données, Taureaux, Spermatozoïdes, Modélisation

Résumé : Dans les élevages la semence bovine est utilisée pour réaliser un grand nombre d'inséminations artificielles afin de renouveler le cheptel bovin et de diffuser des patrimoines génétiques d'intérêt. La présence de taureaux subfertiles parmi les reproducteurs entraîne des pertes économiques pour un grand nombre d'acteurs de l'élevage. Identifier ces animaux subfertiles est donc un enjeu important pour la filière. L'étude des paramètres fonctionnels de la semence ne permet pas d'identifier tous les taureaux subfertiles. De plus, la fertilité mâle est peu héritable et donc difficile à sélectionner à l'aide d'informations génomiques ; si bien qu'il n'existe aujourd'hui pas d'évaluation en routine de la fertilité mâle.

La méthylation de l'ADN est une marque épigénétique jouant un rôle majeur dans la fertilité mâle. En effet, elle subit des remaniements d'ampleur au cours de la différenciation des cellules germinales mâles en spermatozoïdes, ainsi qu'après la fécondation dans le pronoyau mâle et au cours du développement embryonnaire. Elle pourrait donc être le témoin des différences de fertilité pouvant exister entre les taureaux. Le premier objectif de cette thèse a été d'étudier le potentiel du méthylome spermatique comme source de biomarqueurs de fertilité. Pour cela, la semence de 120 taureaux de race Montbéliarde et 38 de race Holstein a été analysée par RRBS (« Reduced Representation Bisulfite Sequencing »). En réalisant des analyses différentielles de méthylation entre des animaux fertiles et subfertiles, des sites différentiellement méthylés (DMC) ont pu être identifiés. Une partie de ces DMC était associée à des gènes impliqués dans la physiologie du spermatozoïde et dans le développement embryonnaire, permettant donc d'émettre des hypothèses biologiques quant à leur lien avec la fertilité. A partir de de ces DMC et en exploitant la méthodologie des forêts aléatoires, il a été possible de construire des modèles permettant de prédire la fertilité avec une précision de 72 à 94% en fonction de la cohorte analysée.

Néanmoins, la méthylation de l'ADN n'est pas le seul facteur important pour la fertilité mâle qui est un phénotype complexe. Prendre en compte plusieurs sources d'informations biologiques pourrait permettre d'augmenter la robustesse des modèles et d'améliorer les connaissances sur les mécanismes moléculaires régulant la fertilité. C'est pourquoi le deuxième objectif de cette thèse a consisté à intégrer les données de méthylation de l'ADN avec des données d'expression des petits ARN non codants et des paramètres fonctionnels de la semence obtenue sur les mêmes échantillons, ainsi qu'avec le génotype des taureaux. Des modèles de prédiction plus performants ont pu être élaborés à partir de 50 à 500 variables issues des différents types d'-omiques, mais jamais des paramètres fonctionnels de la semence qui semblaient peu liées à la fertilité des taureaux. Peu de corrélations ont pu être observées entre les différentes données –omiques, qui semblent plutôt agir de manière complémentaire dans la construction de la fertilité mâle, en conformité avec le statut transcriptionnel particulier du spermatozoïde.

Dans leur ensemble, ces résultats démontrent que l'utilisation de données épigénétiques et génétiques permet de prédire avec une bonne précision la fertilité des taureaux à partir d'un échantillon de semence. Ces résultats, obtenus sur des cohortes de dimensions relativement modestes, peuvent être considérés comme des résultats encourageants, mais nécessitent d'être validés sur des populations plus larges. Néanmoins, ces données offres des perspectives intéressantes pour la compréhension des mécanismes moléculaires de la fertilité mâle, et permettront à court terme d'implémenter un outil d'évaluation de la fertilité des reproducteurs. A plus long terme, l'étude de la transmission de ces variations de l'épigénome spermatique à la descendance pourrait entrainer des avancées majeures dans le domaine de l'élevage.

Title : Identification of biomarkers for bovine semen quality, statistical modelling and biological hypotheses

Keywords : Epigenetic, Data integration, Biomarkers, Bulls, Spermatozoa, Modelling

Abstract : In the bovine industry, bull semen is widely used to carry out a large number of artificial inseminations in order to renew the bovine herds and to disseminate genetics of interest. The presence of subfertile bulls among the breeding stock causes economic losses for a large number of actors in the breeding industry. Identifying these subfertile animals is therefore an important issue for the breeding sector. The study of semen functional parameters does not allow the identification of all subfertile bulls. Furthermore, male fertility displays a weak heritability and is therefore difficult to select using genomic information; therefore, there is currently no routine evaluation of male fertility.

DNA methylation is an epigenetic mark playing a major role in male fertility. Indeed, it undergoes extensive changes during the differentiation of male germ cells into spermatozoa, as well as after fertilization in the male pronucleus and during embryonic development. It could therefore reflect the differences in fertility that may exist between bulls. The first objective of this thesis was to study the potential of the sperm methylome as a source of fertility biomarkers. For this purpose, the semen of 120 Montbéliarde and 38 Holstein bulls was analyzed by RRBS (Reduced Representation Bisulfite Sequencing). By performing differential methylation analyses between fertile and subfertile bulls, differentially methylated sites (DMCs) could be identified. Some of these DMCs were associated with genes involved in sperm physiology and embryonic development, thus allowing us to make biological hypotheses about their link with fertility. From these DMCs and by using the random forest methodology, it was possible to build models that predicted fertility with an accuracy of 72 to 94% depending on the analyzed cohort.

Nevertheless, DNA methylation is not the only important factor for male fertility, which is a complex phenotype. Taking into account several sources of biological information could increase the robustness of prediction models and improve knowledge of the molecular mechanisms regulating fertility. Therefore, the second objective of this thesis was to integrate DNA methylation data with small non-coding RNA expression data and semen functional parameters obtained on the same samples, as well as with the genotypes of the bulls. Prediction models with better performances when compared with DNA methylation only could be developed using 50 to 500 features from the different types of -omics, but never from semen functional parameters , which seemed to have little relationship with the fertility of the bulls. Few correlations could be observed between the different -omics datasets, which rather seem to act in a complementary way in the construction of male fertility, in accordance with the particular transcriptional status of the spermatozoa.

Taken together, these results demonstrate that the use of epigenetic and genetic data allows the prediction of bull fertility with good accuracy from semen samples. These results, obtained on a relatively small cohort, can be regarded as encouraging, but need to be validated on larger populations. Nevertheless, these data offer interesting perspectives for the understanding of the molecular mechanisms underlying male fertility, and will allow in the short term to implement a tool for the evaluation of bull fertility. In the longer term, the analysis of the transmission of these fertility-related epigenetic variations from sperm to the offspring could lead to major advances in the field of breeding.

Remerciements

Tout d'abord je tiens énormément à remercier Hélène Kiefer ainsi que Florence Jaffrezic, pour votre encadrement tout au long de ces trois années. Je suis très content et chanceux de vous avoir eu en tant qu'encadrantes. Plus personnellement, Hélène, je tiens sincèrement à te remercier de ta tolérance vis-à-vis de mon écriture, plus précisément de mes fautes d'orthographes en espérant ne pas avoir trop altérer ta vision à force de lire mes phrases. Plus sérieusement, merci beaucoup, c'est à travers nis différentes discussions que j'ai pu évoluer en tant que scientifique. Tu as toujours trouvé du temps pour m'aider dans mon travail, en témoigne un certain dimanche à Jouy suivi par une nuit blanche pour rendre ce manuscrit dans les temps. Je suis très content que l'on puisse continuer à travailler ensemble dans le cadre du LPA ! Florence, merci beaucoup également. Malgré que je ne sois pas issu d'une formation purement statistique, tu as toujours pris le temps de me faire comprendre les différentes méthodes d'intégration de données et de prendre le temps de répondre à mes questions. Au cours de ces 3 années tu as toujours été positive et bienveillante et donc je suis également très content de pouvoir continuer à collaborer avec toi sur différents projets.

Je tiens également à remercier Laurent Schibler pour ton aide précieuse au cours de ces 3 années. Malgré ton emploi du temps très chargé, tu t'es toujours rendue disponible pour m'épauler dans ce travail de thèse. Tu as été un manager formidable que j'admire beaucoup. Tu m'as laissé une grande liberté pour explorer mes pistes de réflexions, tu m'as fait confiance pour présenter mes résultats devant les adhérents et pour tout ça je voudrais te dire un grand merci.

Je voudrais également remercier Hélène Jammes qui a accepté de m'accueillir dans son équipe de recherche pendant ces 3 années de thèse. Evidemment, je tiens à remercier toutes les personnes constituant cette équipe ; malgré les conditions particulières générées par le Covid-19 qui ne sont pas forcément propice à l'intégration de nouveaux arrivants ou à l'échange ; j'ai été ravie de pouvoir faire ma thèse dans cette équipe accueillante et souriante. Je tiens en particulier à remercier Aurélie Chaulot Talmon pour les bons moments et que l'on a passé ensemble, par contre je ne sais pas si c'est réciproque étant donné que tu m'as fait déménagé de bureau après moins d'un an de thèse ;). Merci également pour les 2-3 banques RRBS que tu as faites (ou 200-300 je ne sais plus ;)) pendant que j'attendais désespérément une réponse pour savoir quand je pourrais commencer ma thèse. Et bien évidemment j'aimerais également remercier chaleureusement toutes les autres personnes qui compose ou qui ont composé cette équipe et qui m'ont beaucoup apporté. Donc merci Anne, Mélanie, Charline, Christine, Karine, Angélique, Lorraine, Véronique, Lotfi, Mélodie, Clara et Ailona.

Je voudrais également remercier beaucoup Eli. Tout d'abord, merci d'avoir accepté de me faire une petite place dans ton petit bureau. J'ai énormément appris grâce à toi que ce soit sur le volet physiologique ou professionnelle de la filière que je ne connaissais pas bien avant de commencer mon travail de thèse. Merci également pour tes qualités humaines, ça a (et ce sera) toujours un plaisir d'échanger avec toi ! Même si aujourd'hui tu as décidé de continuer ta carrière dans le domaine des insectes, j'espère qu'on pourra continuer à se voir le soir quand tu seras de passage vers Jouy.

Je voudrais également remercier énormément Luc Jouneau et Anne Frambourg. J'ai commencé ma thèse juste avant l'été 2019. Autant dire que quelques semaines après mon arrivé l'unité était aussi déserte que des résultats d'alignements uniques de régions répétées en séquençage short read. Néanmoins, deux irréductibles bio-informaticiens sont restés pas mal de temps dans les locaux au cours de cet été. A ce moment, on a pu échanger énormément sur mon projet de recherche, sur la bioinfo et les biostats. Ces moments ont été très importants pour moi et m'ont permis de commencer ma thèse sur de bonnes bases. Après ça vous avez toujours suivi avec beaucoup d'attention mes différents travaux. Pour tout ça je vous dis en grand merci et je suis très content de pouvoir continuer à travailler avec vous à l'avenir.

J'aimerais également beaucoup remercier mes collègues de congrès Sébastien Taussat et Clémentine Escouflaire pour les agréables moments passés lors de l'EEAP, à partir de maintenant je n'accepterai plus de participer à des congrès avec des hôtels moins bien (je pourrais même amener mon peignoir perso maintenant ;)). Clémentine, hâte de pouvoir continuer à tester et parler de jeux de sociétés dans les semaines à venir. Seb je t'aime bien hein, mais autant j'ai la condition physique adéquate pour jouer aux jeux de sociétés, autant pour le rugby il va y avoir un problème. Donc bon à la limite je voudrais bien jouer la troisième mi-temps avec toi, mais compte pas sur moi pour les deux premières ☺.

Je voudrais aussi remercier les différents membres de l'équipe G2B en particulier Chris Hoze, Sébastien Fritz, Didier Boichard, Marie Pierre Sanchez et Mekki Boussaha pour m'avoir aidé tout aux long de ma thèse, sur diverses questions autour de la génétique.

Je tiens également à remercier les membres de mon comité de thèses Patricia Fauque, Julien Chiquet, Carien Capel et Pierre Larraufie pour les discussions pertinentes autour de mon sujet de thèse que l'on a pu avoir au cours de ces 3 années.

Je tiens également à remercier les différents membres de mon jury de thèse d'avoir accepté d'évaluer mon travail au travers de mon manuscrit et de ma soutenance de thèse. Merci au Professeur Joël Drevet de m'avoir fait l'honneur de présider mon jury de thèse. Merci au Dr Julie Cocquet et au

Professeure Sandrine Lagarrigue d'avoir accepté d'être rapporteur de ma thèse et merci au Dr David Makowski d'avoir accepté d'être examinateur de mon travail de thèse.

Je voudrais également à remercier Gildas Mazo et Denis Laloë pour m'avoir aidé à préparer ma soutenance de thèse.

Evidemment, je remercie l'entreprise Eliance de m'avoir fait confiance pour travailler sur ce sujet de thèse et également pour continuer de travailler avec eux après ma thèse. J'ai hâte de pouvoir continuer de travailler avec mes nouvelles collègues du LPA, Chrystelle, Aurélie et Marie-Christine. Je tiens également à remercier Apis-Gene ainsi que l'ANRT pour le financement de ce travail.

Je tiens également à remercier les différentes entreprises partenaires de ce projet : Auriva, l'Awe, Evajura, Evolution et Umotest pour les différents échanges que l'on a pu avoir au cours de ma thèse ainsi que pour nous avoir transmis un grand nombres d'échantillons biologiques.

Merci aux Champions : Andaine, Erwan, Louise, Sébastien, Thibault et pour tous les moments passés ensemble et qui ont permis de me déconnecter de ma thèse.

Merci également à ma Maman et mon frère Naël qui m'ont soutenu au cours de ces 3 ans, au cours de ma soutenance et d'une manière générale depuis des années

Enfin, j'aimerais remercier énormément Nathalie Laforge. Au cours de ces 3 années, tu as été la personne qui a été le plus à mes côtés. Pendant la période de rédaction, alors que j'étais ronchon, stressé et pas très agréable à vivre, tu as toujours été là pour me soutenir et pour m'aider, et c'est en grande partie grâce à toi que j'ai pu tenir et rendre ce travail (presque) dans les temps.

Tables des matières

TABLES DES MATIERES 5
LISTE DES FIGURES 8
LISTE DES ANNEXES 10
LISTE DES ABREVIATIONS 11

INTRODUCTION 13

PREAMBULE : 14
I CONTEXTE GENERAL 16
I.I : L'organisation de la filière a été façonnée par la mise en place de l'IA 16
I.I.I : Introduction et conséquence de l'arrivée de l'IA en France 16
I.I.II : Organisation de la filière 17
I.I.III : Testage sur descendance, évaluation génomique et conséquence sur l'identification des taureaux subfertiles 18
I.II : Les travaux de l'évaluation de la fertilité mâle chez le bovin 20
I.II.I : Indicateurs de fertilité mâle utilisés chez le bovin 20
I.II.II : Peut-on utiliser l'information génomique pour prédire et améliorer la fertilité des mâles ? 21
I.II.III : Peut-on utiliser les paramètres spermatiques pour prédire la fertilité ? 24
I.II.III.I : Le spermatozoïde, une cellule hautement différenciée 24
I.II.III.II : Les altérations fonctionnelles du spermatozoïde 27
I.II.III.III : Les paramètres spermatiques ne permettent pas d'identifier tous les taureaux subfertiles 29
Conclusion 30
II EPIGENETIQUE ET FERTILITE MALE 31
II.I : Les histones 31
II.I.I : Les marques post-traductionnelles des histones 31
II.I.II : L'implication des marques post-traductionnelles d'histones dans la fertilité 33
II.II : La méthylation de l'ADN 34
II.II.I : Machinerie enzymatique en lien avec la méthylation de l'ADN 34
II.II.II : Fonctions biologiques de la méthylation de l'ADN 37
II.II.II.I : Régulation de la transcription 37
II.II.II.II : Répression des éléments répétés 38
II.II.II.III : Inactivation du chromosome X 39
II.II.II.IV : Empreinte parentale 40
II.II.III Reprogrammation de la méthylation de l'ADN : de la différenciation des cellules germinales mâles aux premières étapes du développement 41
II.II.IV : Implication de la méthylation de l'ADN dans la fertilité mâle 46
II.II.V : Implication de la méthylation de l'ADN dans la subfertilité bovine 48
II.III : Les petits ARN non codants 51
II.III.I : Les miRNA 51
II.III.II : Les piRNA 52
II.III.III : Les rsRNA et tsRNA 53
II.III.IV : Dynamique et fonction des petits ARN non codants dans les cellules germinales et la fertilité mâles 53
Conclusion 55
III LES METHODES D'INTEGRATION DE DONNEES 57
III.I : Approches exploratoires 58
III.I.I : L'Analyse Factorielle Multiple 58
III.I.II : Etude des relations entre variables 60
III.I.II.I : L'inférence de réseaux 60
III.II : Prédiction des phénotypes 63
III.II.I : Régression logistique avec pénalité Lasso 63
III.II.II : Forêts aléatoires et Gradient Boosting 64
III.II.II.I : Arbres CART 64

IIII.II.II.II : Forets aléatoires 65
III.II.II.III : Gradient boosting 66
III.II.III Réseaux de neurones 67

RESULTATS 69

OBJECTIFS DE LA THESE ET STRUCTURE DE LA PARTIE RESULTATS 70
COHORTE D'ANIMAUX ET DONNEES TRAITEES AU COURS DE LA THESE 73
Méthylation de l'ADN 76
Les petits ARN non codants 78
Les paramètres spermatiques 79
Les génotypes 80
RESULTATS 81
I Le méthylome spermatique et son utilisation dans la prédiction de fertilité des taureaux. 81
I.I : Prise en compte et traitement des CpG polymorphes 81
I.I.I : Contexte 81
I.I.II : Cohortes utilisées 82
I.I.III : Résultats 82
Conclusion 87
I.II : Analyse du méthylome spermatique en relation avec la fertilité mâle 88
I.II.I : Contexte : 88
I.II.II : Article 1 90
I.II.III : Analyse en race Holstein 91
I.II.IV : Comparaison des résultats obtenus dans les deux races 94
I.III : ETUDES LONGITUDINALES : COHORTES « AGE » ET « DEVIATION » 97
I.III.I : Contexte 97
I.III.II : Analyse de la cohorte Age 98
I.III.III : Analyse de la cohorte « déviation » 99
Conclusion 100
II : Intégration des données pour une prédiction plus fiable de la fertilité 102
II.I : Analyse intégrative en race Montbéliarde 103
II.I.I : Résumé de l'article 2 103
Article 2 104
II.I.II : Résultats préliminaires en inférence de réseaux 105
Problématique 105
Méthode 105
Résultats 107
Discussion 108
II.II : Intégration de données en race Holstein 109
Discussion 114

DISCUSSION GENERALE 117

Cohortes analysées 119
Analyse des données de méthylation 120
Impact des génotypes 121
Fiabilité des résultats de RRBS 124
Impact biologique de variations de la méthylation de l'ADN 126
Application pratique des résultats de la thèse sur le terrain 129
Prédiction de la fertilité au niveau du taureau ou de l'éjaculat 129
Quels types de données -omiques choisir pour prédire la fertilité ? 130
Quelles perspectives pour la prédiction de la fertilité mâle à partir de données épigénétiques ? 133
Héritage du méthylome spermatique à l'embryon 134
Conclusion 137

BIBLIOGRAPHIE 139

ANNEXES 157

COMMUNICATIONS 172

Liste des figures

Figure 1 : Gain génétique de la fertilité femelle pour les vaches Holstein aux Etats-Unis entre 1985 et 2015 19

Figure 2 : Testicule et spermatogénèse 21

Figure 3 : Tractus génitale mâle bovin 23

Figure 4: Schéma d'un spermatozoïde. 23

Figure 5 : Les spermatozoïdes inséminés dans les voies génitales femelles. 24

Figure 6 : La chromatine 28

Figure 7 : La cytosine et la 5' méthyl-cytosine. 31

Figure 8 : Transmission de la méthylation de l'ADN par l'enzyme DNMT1 32

Figure 9 : Schéma illustrant la voie active de déméthylation médiée par les enzymes TET 33

Figure 10 : Régulation de la transcription par la méthylation de l'ADN 34

Figure 11 : Mise en place de l'empreinte parentale 38

Figure 12 : Les deux vagues de reprogrammation au cours du développement 39

Figure 13 : Génération et mécanismes d'actions des miRNA 49

Figure 14 : Deux stratégies permettent aux piRNA de réguler la production de rétrotransposons dans les cellules 50

Figure 15 : Schéma de l'AFM 55

Figure 16 : Illustration de réseaux, et concept de corrélation partielle 57

Figure 17 : Exemple de construction d'une forêt aléatoire avec un jeu de données constitué de 5 individus et de 5 variables 61

Figure 18 : Exemple de construction d'un Gradient Boosting, dans un problème de régression, avec un jeu de données composé de quatre individus de deux variables explicatives (x1 et x2) et d'une variable à prédire (y) 62

Figure 19 : Description des cohortes de taureaux 69

Figure 20: Performances de fertilité des animaux de la cohorte Montbéliarde et Holstein 69

Figure 21 : Les différentes étapes de l'analyse de la méthylation de l'ADN par RRBS 72

Figure 22 : Les paramètres spermatiques 75

Figure 23 : Impact d'un polymorphisme de séquence sur la distribution de la méthylation de l'ADN 79

Figure 24 : Nombre d'allèles CpG au sein d'un CpG polymorphe en fonction du génotype des animaux 80

Figure 25 : Exemple de profils de méthylation avant et après correction par les génotypes 81

Figure 26 : Les incohérences entre niveau de méthylation et génotype sont majoritaires..... 82

Figure 27 : La qualité d'imputation est le facteur expliquant une partie des incohérences entre niveaux de méthylation et génotypes 83

Figure 28 : Les CpG du background ne permettent pas de différencier les animaux fertiles des animaux subfertiles en race Holstein..... 87

Figure 29 : Les DMC sont réparties sur tous les chromosomes 88

Figure 30 : Localisation génomique des DMC 88

Figure 31 : Résultats des analyses d'enrichissement par DAVID..... 89

Figure 32 : Capacité de prédiction du modèle élaboré à partir des DMC spermatiques dans la race Holstein 90

Figure 33 : L'intersection entre les DMC des deux races n'est pas liée au hasard 91

Figure 34 : Etude de la méthylation de l'ADN en fonction de l'âge des animaux..... 94

Figure 35 : Niveau de méthylation sur l'ensemble des CpG du background, en fonction de la cohérence par rapport à la fertilité moyenne des taureaux..... 95

Figure 36 : Inférence de réseaux, montrant que les variables interagissent principalement au sein d'une seule couche d'-omiques..... 103

Figure 37 : Les inférences de réseaux basées sur la méthode des cforest n'identifient que des interactions au sein d'une couche unique d'-omiques 103

Figure 38 : Pré-traitement des données pour l'intégration de données en race Holstein 105

Figure 39 : Résultats de l'AFM en race Holstein..... 106

Figure 40 : Heatmap des variables quantitatives les mieux représentées par le premier axe de l'AFM 107

Figure 41 : Présentation de l'optimisation des hyper-paramètres des modèles..... 107

Figure 42 : Performances des différents modèles en fonction du nombre de variables incluses 108

Figure 43 : Analyse des variables sélectionnées par les trois méthodes 108

Figure 44 : Analyse d'enrichissement fonctionnel 109

Figure 45 : Analyse de l'impact d'un polymorphisme d'un CpG sur le niveau de méthylation d'une DMC 119

Figure 46 : Variation du niveau de méthylation sur un réplicat technique 120

Liste des annexes

Annexe 1 : Taureaux et échantillons de la cohorte longitudinale « Age ».

Annexe 2 Taureaux et échantillons de la cohorte « déviation ».

Annexe 3 : Figure supplémentaire et table supplémentaire de l'article 1.

Annexe 4 : Taureaux constituant le dispositif « Fertilité » en race Holstein.

Annexe 5 : Qualité des séquences et des alignements de la cohorte « Fertilité » en race Holstein.

Annexe 6 : Données supplémentaires de l'article 2.

Annexe 7 : Article : « The epigenome of male germ cells and the programming of phenotypes in cattle ».

Liste des abréviations

5caC	carboxylcytosine
5fC	formylcytosine
5hmC	hydroxymethylcytosine
ACM	Analyse des correspondances Multiples
ACP	Analyse en Composante Principale
ADN	Acide DésoxyriboNucléiques
AFM	Analyse Factorielle Multiple
AGO	Argonaute
ALH	amplitude des mouvements latéraux de la tête
ARN	Acide Ribonucléique
ARNm	ARN messager
AUC	Area Under the Curve
CART	Classification And Regression Tree
CASA	Computer Assisted Semen Analysis
CCA	Analyse des Corrélations Canoniques
CpG	Cytosine - phosphate - Guanine
DAVID	Database for Annotation, Visualisation and Integrated Discovery
DMC	Cytosines Différentiellement Méthylées
DMR	Régions Différentiellement Méthylées
DNMT	DNA methyltransferase
DOHaD	Developmental Origin of Health and Disease
EMP	Entreprise de Mise en Place
ES	Entreprise de Sélection
FIV	Fecondation In Vitro
FSH	Hormone Folicostimulante
GnRH	Gonadotropin-Releasing Hormone
IA	Insémination Artificielle
ICF	Centromeric Instability and Facial anomalies
ICSI	Intra Cytoplsasmic Sperm Injection
IGF2	Insulin Like Growth Factor
jpc	jour post coitum

LASSO Least Absolute Shrinkage and Selection Operator

LH Hormone Lutéinisante

LINE Long Interspersed Nuclear Elements

lncRNA long ARN non codant

MBD Methyl CpG Binding Domain

MH1 Montbeliard Homozygote 1

miRNA micro-ARN

MOFA Multi-Omics Factor Analysis

NOA Azoospermie Non Obstructive

OA Azoospermie Obstructive

PGC Primodial Germ Cells

piRNA ARN intéragissant avec PIWI

PLS Regression des moindres carrés partiels

PMD Partially Methylated Domain

POHaD Paternal Origin of Health and Disease

QTL Quantitative Trait Locus

RISC RNA Induced Silencing Complex

rsRNA fragments dérivés des ARN ribosomiques

SCR Sire Conception Rate

sncRNA petit ARN non codant

SNP Single Nucleotide Polymorphism

STR rectitude du déplacement

TET Ten Eleven Translocation

TNR 56 Taux de Non-Retour en chaleur des femelles inséminées après 56 jours

tsRNA fragments dérivés des ARN de transfert

VAP vitesse de trajectoire moyenne

VCL vitesse de trajectoire curvilinéaire

VSL vitesse de trajectoire en ligne droite

Xist X inactivation specific transcript

ZP3 Zona Pellucida Glycoprotein 3

Introduction

Préambule :

La fertilité mâle est un sujet important dans les filières de l'élevage d'un point de vue économique et organisationnel. Dans les élevages, les taureaux sont des reproducteurs utilisés pour améliorer génétiquement les troupeaux et renouveler le cheptel. En fonction des époques, les taureaux pouvaient avoir plusieurs centaines de milliers de descendants (comme le célèbre Jocko Benne) jusqu'à quelques milliers/centaines de nos jours. Dans les élevages laitiers, la majorité des fécondations se font par Insémination Artificielle (IA) (détail dans l'introduction). Cette pratique suppose la sélection et l'élevage d'un taureau, la production et la commercialisation de sa semence et la mobilisation d'un technicien d'insémination dans un élevage. Le succès de l'ensemble de ces opérations est nécessaire afin d'aboutir à la naissance d'un veau et de débuter les lactations des vaches inséminées. La réussite d'une IA, c'est à dire l'obtention d'une gestation et la naissance d'un veau, n'est pas garantie et chaque IA n'étant pas fécondante entraîne des pertes économiques pour différents acteurs de la filière. Ainsi, des taureaux subfertiles, par définition ayant une fertilité quantitativement plus faible que des taureaux fertiles, vont en espérance avoir moins de succès à l'IA que leurs homologues fertiles. De fait, la présence de taureaux subfertiles dans les élevages a un impact négatif sur la filière.

Pour résoudre ce problème, des études ont été menées en analysant la génétique des taureaux ainsi que la qualité fonctionnelle des spermatozoïdes des taureaux, dans le but d'identifier des biomarqueurs précoces de la fertilité mâle bovine. Cependant, ces deux types de paramètres ne permettent pas d'identifier tous les taureaux subfertiles. Ainsi, de nouveaux marqueurs sont nécessaires dans le but d'identifier les taureaux subfertiles.

L'épigénétique pourrait permettre de mieux prédire la fertilité mâle. Il a en effet été montré que l'épigénome spermatique était essentiel pour la différenciation des gamètes mâles, pour la spermatogénèse ainsi que le développement embryonnaire, ce qui en fait une potentielle source de biomarqueurs de fertilité mâle. Ces faits ont motivé la création de SeQuaMol en 2014, un laboratoire commun unissant INRAE et ALLICE, à l'initiative d'H. Jammes et L. Schibler, visant à identifier des

biomarqueurs de la qualité de la semence à partir de données moléculaires et en particulier épigénétiques.

Ce projet de thèse, directement lié aux problématiques de SeQuaMol, s'inscrit dans un cadre pluridisciplinaire à l'interface entre la biologie et les statistiques. Les objectifs sont multiples et incluent l'identification de biomarqueurs de la qualité de semence, la création de modèles de prédiction de la fertilité, la réalisation d'analyses intégratives et l'élaboration d'hypothèses biologiques. Les résultats apportés par ces travaux, en combinaison avec ceux déjà obtenus dans le cadre du projet SeQuaMol, ont pour objectif de développer à terme des outils utilisables sur le terrain dans le but de prédire la fertilité des taureaux. Cette thèse a été menée dans le cadre d'une convention CIFRE (co-financement APIS-GENE et ANRT), au cours de laquelle j'étais employé par ALLICE et détaché à INRAE.

L'introduction ci-dessous aura pour but, dans un premier temps, d'expliquer le contexte général de la filière en France ainsi que les besoins d'une évaluation de la fertilité mâle. Dans un second temps, l'importance des marques épigénétiques du spermatozoïde dans la fertilité mâle bovine sera abordée, avec une emphase sur la méthylation de l'ADN, marque épigénétique principalement étudiée au cours de ce travail. Enfin la dernière partie aura pour objectif de présenter les différents types de méthodes statistiques utilisées au cours de ce travail.

I Contexte général

I.I : L'organisation de la filière a été façonnée par la mise en place de l'IA

I.I.I : Introduction et conséquence de l'arrivée de l'IA en France

Dans le contexte d'après-seconde guerre mondiale, l'agriculture a subi de profonds changements dans le but de répondre aux besoins nutritionnels de la société. Ces profonds changements ont été médiés par des innovations technologiques et organisationnelles importantes. Un de ces changements a été la mise en place de l'IA.

Historiquement, la première IA documentée a été conduite par Lazzaro Spallanzani au 18ème siècle, et a été effectuée chez des chiens en prenant le sperme d'un mâle et en l'inséminant dans les voies génitales de femelles ce qui a eu pour conséquence la naissance de trois chiots. Cette expérience a pu montrer que les spermatozoïdes étaient nécessaires à la fécondation. Cela montrait de plus qu'il y avait besoin d'un mâle et d'une femelle pour produire une descendance, mais que le contact physique entre ces derniers n'était pas nécessaire. Fort de ces observations, et des premières applications de l'IA dans des contextes zootechniques par le russe Ilya Ivanovich Ivanoff au début de 20ème siècle (Lonergan, 2018), les français Robert Cassou et Martial Laplaud ont optimisé ce processus dans les années 1940-1950 afin de le rendre viable pour une application industrielle, en créant la « paillette Cassou » (Livre : De la paillette à l'ère du génome ; 70 ans d'aventure humaine). En effet, avant la mise au point de ce support, les éjaculats des taureaux étaient contenus dans des ampoules de verres. Ces supports prenaient de la place et permettaient une conservation des éjaculats de taureaux entre 24 et 48 heures au maximum ce qui rendait difficile leur diffusion. La « paillette Cassou » est individualisée, prend peu de place et peut-être cryoconservée afin de conserver le matériel biologique dans de l'azote liquide pendant de longues périodes, permettant d'en faciliter le transport et la diffusion. Chaque paillette permet de contenir une partie de la semence d'un taureau, afin de réaliser une IA. Ainsi, à partir d'un éjaculat, il était possible de produire un grand nombre de paillettes pour subvenir aux

besoins de reproduction des troupeaux. Cette paillette pouvait ensuite être décongelée et utilisée par un technicien d'insémination afin d'inséminer une vache.

Dans un contexte d'élevage, ces innovations ont eu des conséquences logistiques importantes, qui ont placé les taureaux au centre du renouvellement des cheptels et de l'amélioration génétique des bovins. En effet, les caractères d'intérêts agronomiques sont en général des caractères quantitatifs reposant en partie sur une composante génétique. Ainsi, en sélectionnant et diffusant des taureaux de haute valeur génétique pour un ensemble de caractères d'intérêts agronomiques, la filière bovine s'assurait d'obtenir des vaches produisant de plus en plus, réalisant ainsi une amélioration génétique des troupeaux au fur et à mesure des générations. Cette façon de procéder a été en partie permise grâce à l'IA qui a diffusé des taureaux de haute valeur génétique dans de nombreux élevages bovins.

I.I.II : Organisation de la filière

Il existe un grand nombre d'acteurs impliqués dans la filière, ayant tous des fonctions dans la pérennité de ce système agricole. Parmi ces différents acteurs on peut en discerner trois qui sont directement impliqués dans la production, la diffusion et l'utilisation de la semence bovine :

(i) Les entreprises de sélection (ES) qui ont pour rôle de sélectionner, d'élever, de produire et de conditionner les paillettes de semence de taureaux de haute valeur génétique

(ii) Les entreprises de mise en place (EMP) qui sont chargées de rendre accessible l'IA auprès des éleveurs, en employant des techniciens d'insémination et en déployant les moyens logistiques nécessaires

(iii) Enfin, l'éleveur qui choisit sur catalogue et achète le taureau avec lequel il souhaite réaliser une ou plusieurs IA dans son exploitation.

I.I.III : Testage sur descendance, évaluation génomique et conséquence sur l'identification des taureaux subfertiles

Dans le but d'identifier les taureaux de haute valeur génétique et de diffuser leur semence, des méthodes d'évaluation ont dû être mises en place. Le « testage sur descendance » a été la méthode d'identification des taureaux de haute valeur génétique depuis l'utilisation de l'IA jusqu'au début des années 2010. Elle avait un double avantage, car elle permettait à la fois d'évaluer avec précision la valeur génétique d'un animal et en même temps d'apprécier sa fertilité. Elle consistait à utiliser des taureaux candidats à la sélection (souvent issus d'une ascendance elle-même de haute valeur génétique) et à les évaluer en fonction des performances de leurs filles. A l'âge de procréer, ces taureaux étaient amenés dans des centres de sélection où leur semence était récoltée. Elle était ensuite diffusée par IA dans différents troupeaux et sur différentes femelles. Les veaux femelles issus de cette fécondation étaient ensuite évalués sur divers caractères, comme ceux liés à la production laitière par exemple (volume de lait journalier, taux de protéines et de matières grasses). Si en moyenne sur toute la descendance du taureau le caractère analysé était supérieur à la population générale, alors la génétique du taureau était favorable pour le caractère analysé. De cette façon, un indicateur quantitatif de la génétique du taureau pouvait être déduit afin d'identifier les meilleurs candidats. Seuls les taureaux avec la plus haute valeur génétique étaient ensuite diffusés sur le territoire national afin de maximiser le progrès génétique. Afin d'obtenir une valeur fiable de la valeur génétique des animaux, il fallait donc que chaque taureau ait une descendance suffisante, et donc suffisamment d'IA fécondantes. Ainsi, la fertilité de ces taureaux était en même temps évaluée ce qui permettait d'écarter les taureaux ayant de mauvaises performances de fertilité.

Cependant, le testage sur descendance, bien qu'efficace, avait plusieurs désavantages. En effet, cette évaluation était longue car elle nécessitait que les taureaux soient pubères, qu'ils produisent des doses de semence, qu'elles soient utilisées pour inséminer des vaches et que les premières performances de leurs filles soient mesurées ; cela engendrait des coûts d'évaluations importants et un intervalle de

génération important également. En effet, environ 5 ans étaient nécessaires pour identifier des reproducteurs ayant une bonne génétique, une période pendant laquelle le taureau était entretenu mais non exploité. De plus, les taureaux candidats à la sélection étaient présélectionnés par rapport à leur ascendance ce qui réduisait le nombre de candidats et donc la diversité génétique des potentiels reproducteurs.

C'est pourquoi dans la plupart des races laitières en France vers la fin des années 2000, l'évaluation génomiques des reproducteurs a été mise en place (Boichard *et al.*, 2012), et s'est depuis généralisée. Cette méthode d'évaluation ne nécessite plus d'attendre que les taureaux aient des descendants pour évaluer leur valeur génétique, mais se base uniquement sur le génotype des taureaux. Cela est possible grâce à la mise en place d'une population de référence qui est phénotypée (pour les caractères d'intérêts) et génotypée. Cette population permet d'établir une relation entre le phénotype d'un animal et son génotype. Ainsi en ne connaissant que le génotype d'un candidat à la sélection, on peut estimer sa valeur génétique avec une précision quasi similaire au testage sur descendance. Cette méthode de sélection a été adoptée par la filière, car elle présente plusieurs avantages : évaluation précoce des candidats, peu coûteuse, intervalle de génération réduit, inclusion de plus de candidats à la sélection et donc réduction théorique de la consanguinité. Aujourd'hui avec un recul d'une dizaine d'années, cette méthode a largement fait ses preuves et a presque entièrement remplacé le testage sur descendance.

Cependant, malgré ces qualités indéniables, cette méthode qui se passe de la descendance des taureaux ne permet plus d'évaluer leur fertilité. Ainsi, les taureaux ayant une génétique intéressante sont directement utilisés pour produire des paillettes commercialisables. Les centres de production de semence hébergent ainsi des taureaux de fertilité hétérogène –des taureaux très fertiles jusqu'aux taureaux subfertiles. Comme expliqué dans le préambule, l'utilisation d'animaux subfertiles peut avoir des conséquences négatives d'un point de vue économique et organisationnel pour différents acteurs de la filière. Ces différents acteurs sont directement dépendants de la bonne fertilité des taureaux. Les

ES élevant des taureaux subfertiles, auront du mal à en rentabiliser la semence. Pour les EMP et les techniciens d'insémination employés, de nombreux déplacement devront être assurés dans les élevages, qui sont en zone rurale nécessitant de long trajets. Chaque IA non fécondante nécessite en effet un nouveau déplacement dans l'élevage et l'utilisation d'une nouvelle dose de semence. Enfin les éleveurs qui achètent la paillette sont dépendants de la réussite de l'IA pour initier les lactations des vaches ainsi que pour renouveler leur cheptel.

Chaque acteur individuellement est donc dépendant de la fertilité mâle, démontrant son importance d'un point de vue économique mais également organisationnel pour les différents acteurs de la filière.

I.II : Les travaux de l'évaluation de la fertilité mâle chez le bovin

Devant l'absence d'évaluation permettant précocement de caractériser la fertilité mâle, il y a eu la nécessité d'identifier des indicateurs biologiques adéquats. La génétique et les paramètres spermatiques de la semence, en particulier, ont été exploités dans le but de prédire la fertilité des taureaux. Cependant, avant de décrire plus en détail ces travaux, il convient tout d'abord d'apporter des précisions sur le terme « fertilité » dans un contexte d'élevage.

I.II.I : Indicateurs de fertilité mâle utilisés chez le bovin

Chez le bovin, la fertilité n'est pas définie comme une donnée qualitative comme en clinique humaine, mais comme une donnée quantitative. Cela est possible car la semence de taureaux est utilisée pour réaliser beaucoup d'IA, permettant de quantifier le nombre de fécondations qui ont abouti à la naissance d'un veau. En réalité, le taux de conception, c'est-à-dire si l'IA a pu aboutir à une naissance, n'est pas beaucoup utilisé en France (mais très utilisé aux Etats-Unis par exemple). On lui préfère un autre indicateur plus précoce qui lui est corrélé et qui est le Taux de Non-Retour en chaleur des femelles inséminées après 56 jours (TNR 56). Il est calculé en observant les chaleurs d'une vache ou d'une génisse après avoir réalisé une insémination. Si la vache a manifesté des chaleurs dans les 56 jours après une IA, cela signifie qu'elle est revenue en cycle et par conséquent, que l'IA n'a pas été

« fécondante ». En revanche, si elle n'est pas revenue en chaleur dans cet intervalle de temps, alors l'IA est considérée comme fécondante. Par convention quand une IA n'a pas fécondante on lui attribue la valeur 0, et la valeur 1 dans le cas contraire. Ainsi, en réalisant une moyenne des résultats obtenus sur l'ensemble des IA d'un taureau au cours de sa carrière, on obtient un indicateur exprimé en pourcentage qui quantifie sa fertilité sur le terrain et que l'on appellera le TNR brut. Plus la valeur de ce TNR est proche de 100% plus la fertilité du taureau est élevée et à l'inverse, plus elle est proche de 0% plus elle est faible.

Cependant, cet indicateur est fortement biaisé et on ne peut pas directement lier la valeur du TNR brut à la fertilité du taureau. En effet, la fertilité du taureau n'est pas le seul facteur de variation qui explique la réussite d'une IA, il en existe beaucoup d'autres au-delà de la fertilité de la femelle inséminée, qui est évidemment primordiale pour le succès de l'IA. C'est pour cela que cette valeur brute est corrigée d'un grand nombre de facteurs de variation identifiés comme étant significatifs (Barbat *et al.*, 2010) : année d'insémination, mois d'insémination, jour de la semaine, technicien d'insémination, type de semence (sexée, c'est-à-dire porteuse d'un seul type de chromosomes sexuels, ou conventionnelle), parité de la vache inséminée, intervalle depuis le dernier de vêlage, génétique et environnement permanent de la vache inséminée. A l'issue de cette correction, on obtient un TNR dit corrigé qui reflète les variations de la réussite à l'IA uniquement imputables au taureau. C'est à l'aide de ce TNR corrigé que l'on peut donc comparer les performances de fertilité des taureaux. Plus généralement, c'est en se basant sur un indicateur corrigé (du type TNR ou SCR, « Sire Conception Rate ») que des études ont tenté de mettre en évidence l'association entre ce caractère et des facteurs biologiques.

I.II.II : Peut-on utiliser l'information génomique pour prédire et améliorer la fertilité des mâles ?

La génétique des animaux a rapidement été analysée pour prédire la fertilité mâle. Il existe un triple avantage à identifier des marqueurs génétiques de fertilité.

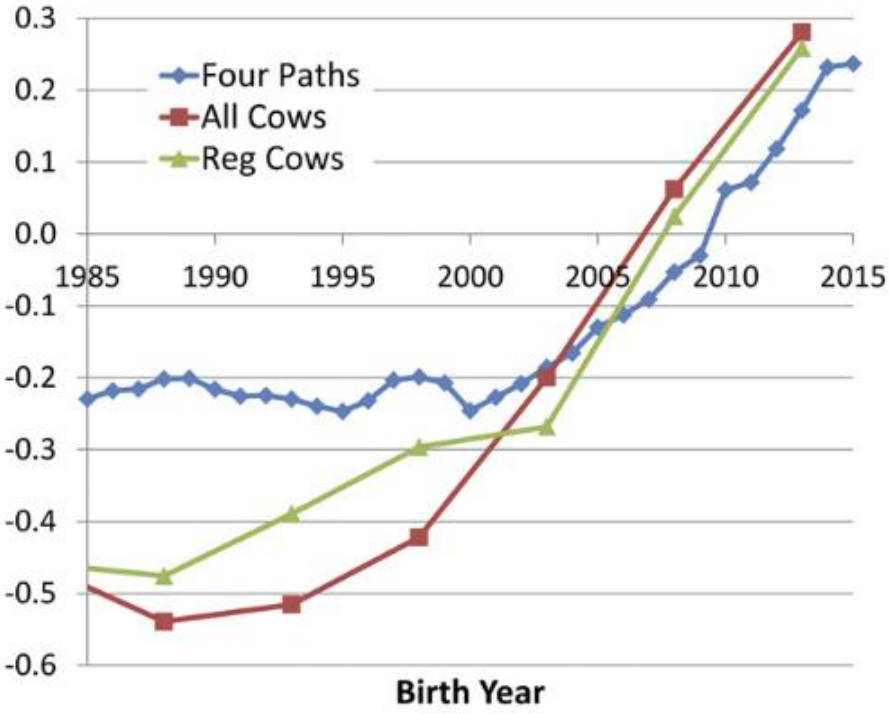


Figure 1 : Gain génétique de la fertilité femelle pour les vaches Holstein aux Etats-Unis entre 1985 et 2015 (tiré de Garcia-Ruiz et al . 2016)

(i) Le premier est que l'on pourrait de manière très précoce (rapidement après la naissance, voire même quelques jours après une fécondation *in vitro*) estimer la fertilité future de l'animal.

(ii) Le deuxième est qu'en mettant la fertilité mâle au cœur des objectifs de sélection, on pourrait potentiellement accroître la fertilité mâle au fur et à mesure des générations.

(iii) Le troisième est que la sélection génomique est déjà largement maîtrisée et mise en place dans l'industrie bovine. Tous les taureaux d'IA sont génotypés ; ainsi, il n'y aurait pas besoin d'effectuer de changements technologiques ou logistiques.

L'héritabilité d'un caractère se définie comme la variance phénotypique qui est imputable à la génétique. C'est une valeur qui est comprise entre 0 et 1 ($h^2 = Vg/Vp$; avec h^2 l'héritabilité, Vg la variance génétique additive et Vp la variance phénotypique) ; des valeurs aux alentours de 0 signifient que la variabilité d'un phénotype n'est que faiblement due à la génétique (elle peut être par exemple due aux impacts environnementaux), et à l'inverse plus l'on se rapproche de 1 plus la variance phénotypique est expliquée par la génétique. Dans le cadre de la fertilité mâle, l'héritabilité du TNR est aux alentours de 0,001 en fonction des études (Berry *et al.,* 2014). Cela montre que la variance du TNR n'est que faiblement due à la génétique de l'animal, ce qui a pour conséquence une grande imprécision de prédiction de ce phénotype à partir du génotype. Cependant, étant donné que l'héritabilité n'est pas nulle, il y a malgré tout une composante génétique de la fertilité mâle. En effet plusieurs Quantitative Trait Loci (QTL) –loci de caractères quantitatifs- sont décrits dans les deux revues suivantes (Fortes *et al.*, 2013; Taylor *et al.,* 2018). De plus, une faible héritabilité ne signifie pas qu'il ne peut pas y avoir de progrès génétique, comme cela a pu être montré en fertilité femelle. En effet, la fertilité femelle aux Etats Unis est mesurée à l'aide d'un indicateur appelé le « Daughter Pregnancy Rate » qui a une héritabilité de 0,04 ce qui également faible. Comme on peut le voir dans la Figure 1, malgré sa faible héritabilité, ce caractère s'est amélioré au fur et à mesure des années montrant un gain de fertilité chez les femelles (García-Ruiz *et al.*, 2016).

En plus de ces aspects, d'autres éléments ont pu être associés à des déficits de fertilité et c'est le cas de certaines régions homozygotes récessives létales. Ces régions, qui n'ont pas d'incidence si elles sont portées à l'état hétérozygote, aboutissent à une létalité précoce à l'état homozygote. Une étude a été réalisée dans les trois grandes races laitières françaises permettant d'identifier un nombre important (18 en race Holstein, 11 en race Montbéliarde et 6 en race Normande) de ces régions récessives létales (Fritz *et al.*, 2013). Du fait de l'absence de phénotype à l'état hétérozygote, un grand nombre de porteurs sains existe dans les élevages. C'est le cas par exemple de la région MH1 en race Montbéliarde qui est portée à l'état hétérozygote par 9% des individus. La découverte des haplotypes récessifs létaux a ainsi permis à la filière bovine française de conduire des accouplements raisonnés en évitant des fécondations entre individus porteurs sains.

Cependant, bien que la génétique constitue une source de marqueurs d'intérêt pour la fertilité mâle chez les bovins, il n'existe pas encore d'évaluation génomique de la fertilité mâle en France, ce qui revient à dire qu'il n'est actuellement pas possible de prédire la fertilité d'un taureau à partir de son génotype. Cela a deux conséquences : basé sur le génotype d'un animal, on ne peut pas connaître sa fertilité ; il n'y a pas de progrès génétique direct de la fertilité mâle dans les élevages. Ici, le mot « direct » est utilisé, car il peut néanmoins y avoir une sélection indirecte de la fertilité mâle. En effet, les paramètres spermatiques, qui permettent de mesurer la qualité de la semence (voir chapitre ci-dessous), et les caractères liés aux capacités reproductives, sont caractérisés par une héritabilité moyenne (de 0, 05 pour la motilité spermatique à 0, 4 pour la circonférence scrotales *(Berry et al.,* 2014)). Comme seuls les reproducteurs capables de produire de la semence de bonne qualité sont diffusés sur le territoire, et que ce caractère repose en partie sur une composante génétique, les taureaux ayant une génétique défavorable pour la production de semence de bonne qualité sont indirectement et progressivement supprimés.

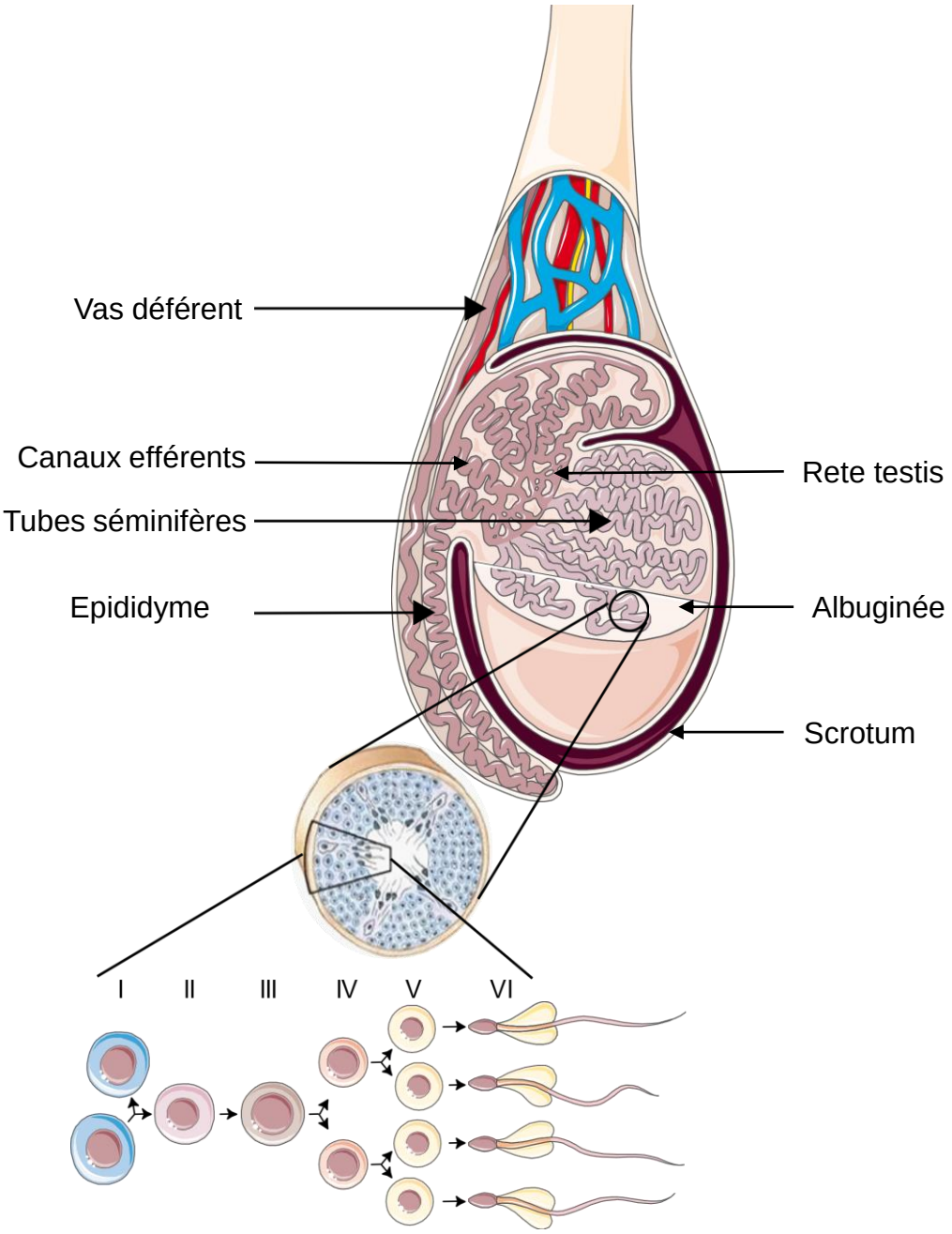


Figure 2 : Testicule et spermatogénèse. La spermatogénèse prend place dans le testicule bovin au niveau des tubes séminifères. Les spermatogonies de type A (I) vont permettre la formation de spermatogonies de type B (II), et assurent leur auto-renouvèlement. Les spermatogonies de types B se différencient en spermatocyte de type I (III) au moment de la première division de méiose. Les spermatocytes de type II (IV) entrent en deuxième division de méiose et à l'issue de ces deux étapes, 4 spermatides rondes seront formées (V). Au cours de la spermiogénèse, ces spermatides vont subir de multiples changements permettant l'obtention de spermatides allongées (VI). Au cours de la spermiation, les spermatides allongées sont larguées dans la lumière des tubes séminifères.

I.II.III : Peut-on utiliser les paramètres spermatiques pour prédire la fertilité ?

Un autre facteur intéressant à étudier quand il s'agit de fertilité mâle est le spermatozoïde, cellule essentielle pour permettre la fécondation et le développement d'une descendance dans un environnement physiologique. Certains paramètres fonctionnels de la semence ont en effet fait l'objet d'intenses efforts de recherche & développement et de standardisation, et sont mesurés en routine lors des procédures de production de semence bovine ou en clinique humaine. De plus, l'utilisation de ces paramètres dans le cadre de l'étude de la fertilité mâle continue de mobiliser la communauté scientifique. Il semble donc légitime de consacrer un chapitre de cette thèse aux altérations des paramètres spermatiques observées dans des cas d'infertilité ou de subfertilité. Cependant, avant d'expliquer quels sont les éléments pouvant affecter la différenciation et la fonction des spermatozoïdes, et donc nuire à la fertilité, il convient de rappeler les différentes étapes permettant d'obtenir un spermatozoïde fécondant après la puberté, en particulier chez le taureau. Cette partie aura donc pour but de décrire la spermatogénèse et la maturation post-testiculaire des spermatozoïdes jusqu'à la fécondation, pour montrer les éléments nécessaires à l'obtention d'un spermatozoïde fécondant puis de présenter les études réalisées dans le cadre de la fertilité bovine.

I.II.III.I : Le spermatozoïde, une cellule hautement différenciée

La spermatogénèse :

Sous l'impulsion de la GnRH (Gonadotropin-Releasing Hormone), la production des hormones LH (Hormone Lutéinisante) et FSH (Hormone Folicostimulante) a lieu, permettant aux taureaux d'entrer dans une phase péri-pubertaire. Cette phase est marquée par le début de la spermatogénèse, processus permettant d'assurer la production de spermatozoïdes encore non matures, mais dont le patrimoine génétique est établi et la morphologie reconnaissable. Ce processus prend place dans le testicule au sein de tubes séminifère (Figure 2). Il est découpé en trois étapes différentes : la spermatocytogénèse, la méiose et la spermiogénèse (Roosen-Runge, 1962; Staub and Johnson, 2018).

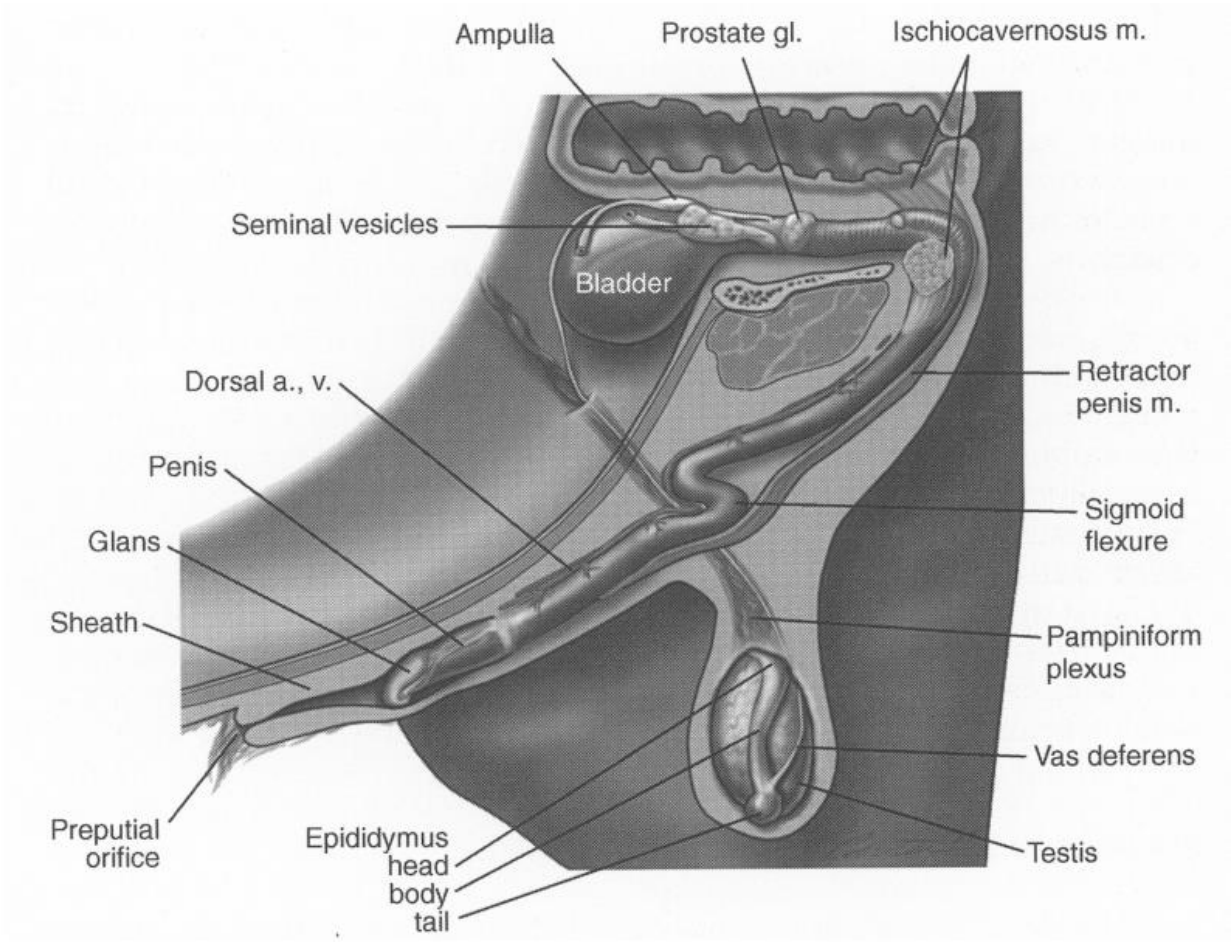


Figure 3 : Tractus génitale mâle bovin. Schéma représentant le tractus génital d'un taureau (Jack C. Whittier 1993).

La spermatocytogénèse comprend de nombreuses divisions cellulaires et permet à la fois d'assurer le renouvellement permanent des spermatogonies, constituant un stock de cellules à partir desquelles la spermatogenèse peut être maintenue tout au long de la vie (spermatogonies de type A) et de produire des cellules entrant en spermatogenèse (spermatogonies de type B). Les spermatogonies de type B, après des étapes de prolifération mitotiques, vont se différencier en spermatocyte au moment de la méiose.

Cette étape permet de générer quatre cellules haploïdes à n chromosomes à partir d'une cellule de spermatogonie B. La première division de méiose est qualifiée de réductionnelle, car le nombre de chromosomes est divisé par deux après réplication de l'ADN. Avant de subir cette division réductionnelle, le spermatocyte primaire contient donc transitoirement deux fois plus de matériel génétique (4n) qu'une cellule somatique (2n). La deuxième division de méiose sera rapidement enclenchée par les spermatocytes secondaires (2n) après la fin de la première et sera qualifiée d'équationnelle, car cette fois-ci le même nombre de chromosomes est conservé, mais les chromatides sœurs sont séparées au niveau de chaque chromosome (1n).

La spermiogénèse, dernière étape de la spermatogénèse, permet la transition des spermatides rondes nouvellement formées en spermatozoïdes. Ce processus est caractérisé par la maturation des spermatides rondes sans aucune division cellulaire. Il comprend trois principaux changements :

(i) La formation de l'acrosome. Cette organelle indispensable à la fécondation (voir plus loin) est formée à partir de l'appareil de Golgi et des vésicules pré-acrosomale (Foster and Gerton, 2016).

(ii) La compaction du noyau qui s'accompagne de changements drastiques de la chromatine. En effet, 85 à 99% des histones (en fonction des espèces) sont remplacés par des protamines, protéines nucléaires permettant une plus grande compaction de l'ADN via des structures toroïdales (Rathke *et al.*, 2014).

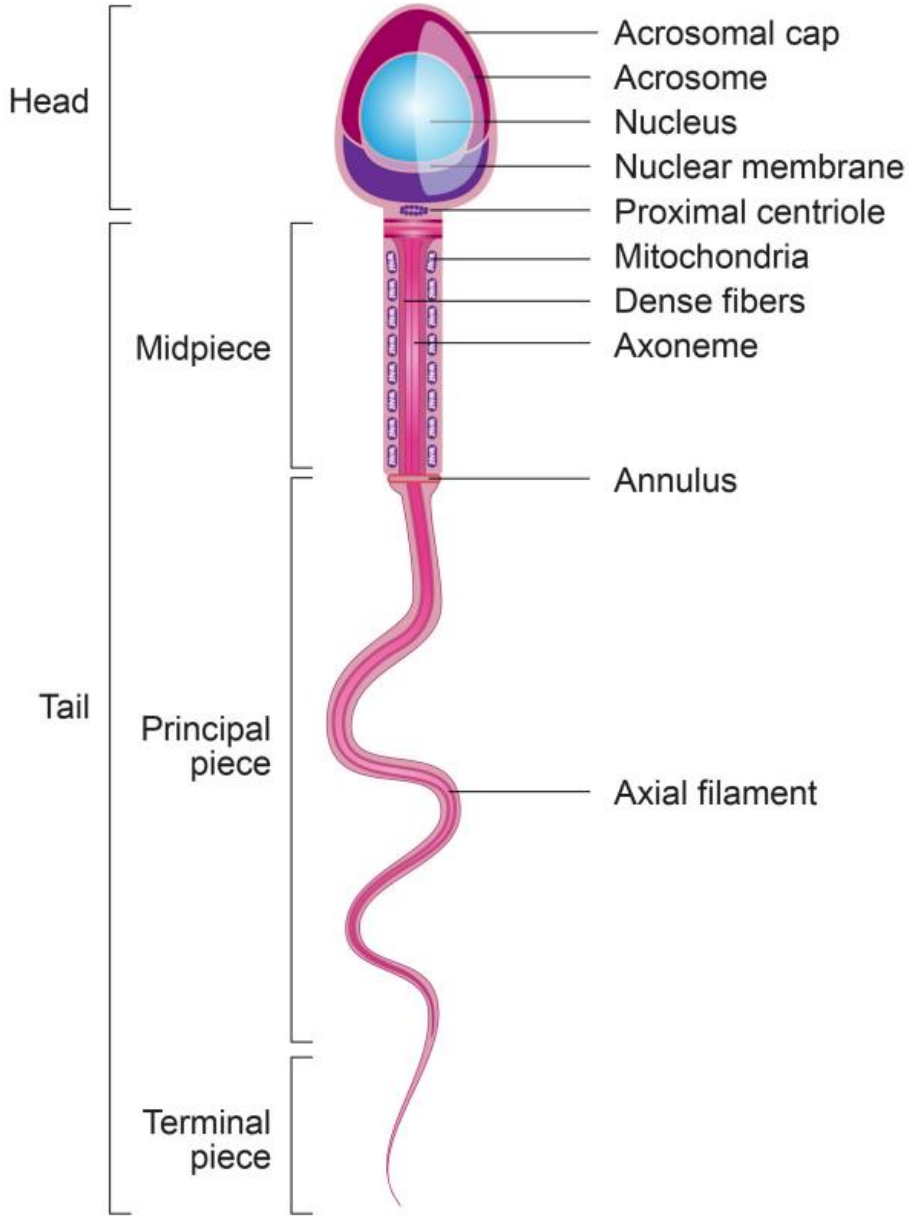


Figure 4: Schéma d'un spermatozoïde. Les spermatozoïdes sont constitués de 3 parties, la tête incluant l'acrosome et le noyau, la pièce intermédiaire contenant les mitochondries et le flagelle permettant la motilité du spermatozoïde (tiré de Alves, Celeghini, and Belleannée 2020)

(iii) La formation du flagelle. Cette structure va prendre forme tout au long de la spermiogénèse via le développement de la taille des microtubules. En parallèle, les mitochondries sont relocalisées dans la pièce intermédiaire permettant de subvenir aux besoins énergétiques du spermatozoïde quand sa motilité sera acquise (Lehti and Sironen, 2017).

A l'issue de ces étapes de spermiogénèse, un spermatozoïde immature et immotile est relargué dans la lumière des tubes séminifères via un processus appelé la spermiation qui conclut la spermatogénèse. Chez le taureau ce processus dure 61 jours, avec 21 jours pour la spermatocytogénèse, 23 jours pour la méiose et 17 jours pour la spermiogénèse. Cependant le spermatozoïde nouvellement formé n'est pas encore fécondant, et pour le devenir, il doit subir d'autres phases de maturation dans le tractus génital mâle (Figure 3).

La maturation :

A la sortie du testicule, les spermatozoïdes migrent jusqu'à l'épididyme. Cet organe contient trois parties distincte : la tête, le corps et la queue. Il est long de plusieurs dizaines de mètres chez le bovin et il est nécessaire pour le transport, la maturation et le stockage des spermatozoïdes avant l'éjaculation. Les spermatozoïdes entrant dans l'épididyme sont immotiles et non fécondants ; au cours de leur transit au travers de cet organe, plusieurs modifications fonctionnelles vont avoir lieu et vont permettre de leur conférer leur motilité et leur pouvoir fécondant (Figure 4) (Sullivan and Mieusset, 2016). Au moment de l'éjaculation, les spermatozoïdes rencontrent le plasma séminal issu des vésicules séminales, une étape qui va permettre d'améliorer leur fécondance (Okamura *et al.*, 1985; Leahy and De Graaf, 2012).

Les voies génitales femelles et la fécondation :

Dans le domaine de l'élevage bovin laitier, la plupart des fécondations sont réalisées par IA. Dans le cas des IA, il n'y pas de rencontre physique entre le taureau et la vache à inséminer. Le sperme est introduit dans la vache à l'aide d'un pistolet à insémination, directement au niveau du corps de l'utérus

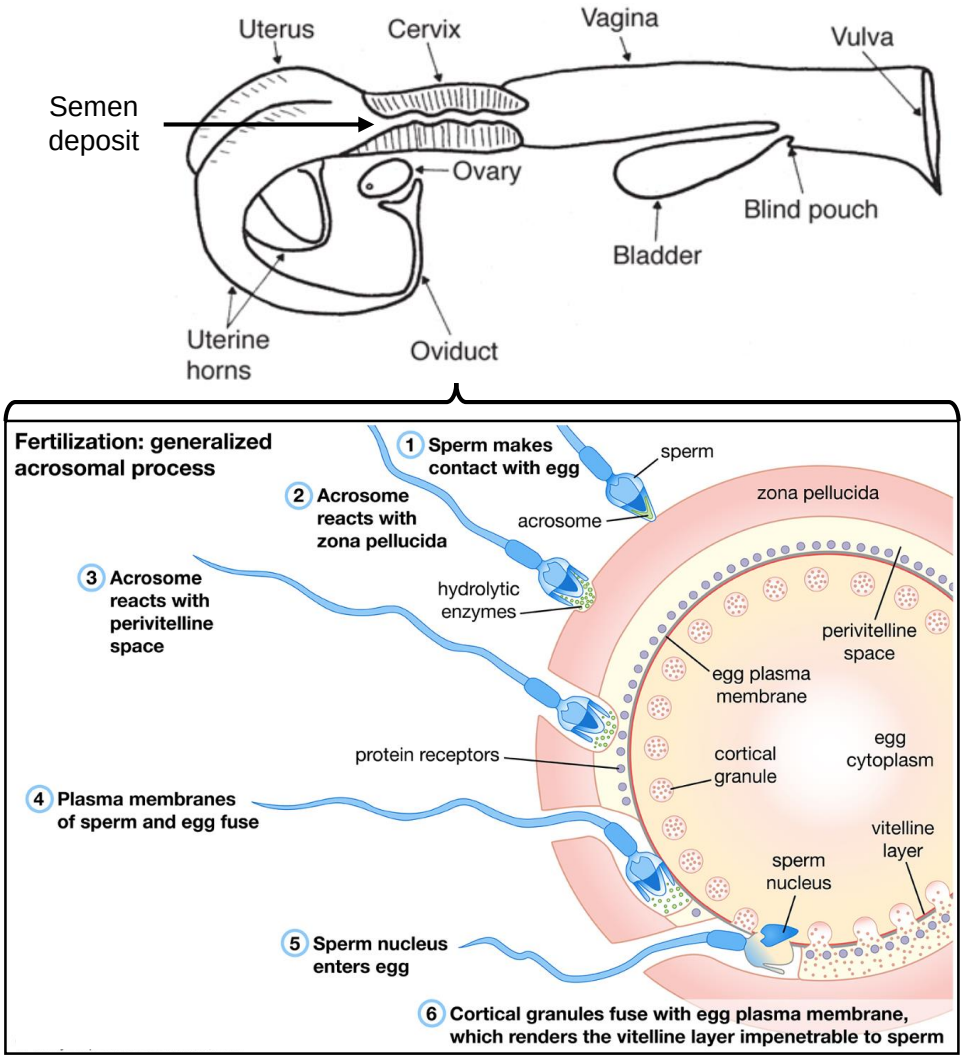


Figure 5 : Les spermatozoïdes inséminées dans les voies génitales femelles. Au moment de l'insémination artificielle, les spermatozoïdes sont déposés dans le corps utérin de la vache inséminée. A l'aide des contractions musculaires de l'utérus les spermatozoïdes vont pouvoir remonter les cornes utérines pour s'établir au niveau de la jonction utéro-tubulaire. Une partie des spermatozoïdes va remonter l'oviducte jusqu'à l'ampoule tubulaire. Au moment de l'interaction d'un spermatozoïde avec la zone pellucide, la réaction acrosomique va avoir lieu, lui permettant de la traverser. Le spermatozoïde va ensuite fusionner avec l'oocyte, permettant l'initiation du développement embryonnaire (schémas tiré de Fayrer-Hosken 1997 et Encyclopaedia Britanica)

(Figure 5). Le vagin et le col de l'utérus sont en effet des milieux hostiles pour les spermatozoïdes, et seulement une faible proportion d'entre eux franchissent cette barrière naturelle (Suarez and Pacey, 2006). Ainsi, en déposant directement les spermatozoïdes dans l'utérus, il y a moins de pertes, les paillettes inséminées peuvent donc contenir moins de spermatozoïdes. Les spermatozoïdes ainsi déposés, migrent ensuite jusqu'à la jonction utéro tubulaire à l'aide des contractions musculaires de l'utérus (Hawk, 1987). Au cours de son trajet dans le tractus génital femelle, le spermatozoïde est confronté à différents agents qui modifient sa physiologie pour promouvoir la capacitation -une série d'événements permettant au spermatozoïde d'être hyperactivé, ce qui se traduit par une augmentation de sa mobilité permettant d'atteindre l'oocyte, et de le préparer à la réaction acrosomique (De Jonge, 2005). Une fois arrivé au niveau de l'ampoule tubulaire, le spermatozoïde entre en contact avec la zone pellucide en interagissant avec ses récepteurs de la famille des ZP dont ZP3 (Zona Pellucida Glycoprotein 3) (Florman and Wassarman, 1985). Au niveau du spermatozoïde, cette interaction a pour conséquence d'enclencher la réaction acrosomique. Cette réaction déclenche le relargage du contenue de l'acrosome. Certains des composants de l'acrosome sont des hyaluronidases, qui vont permettre de faciliter le passage du spermatozoïde jusqu'à l'oocyte. Une autre protéine larguée par l'acrosome est IZUMO1, indispensable pour la fusion du spermatozoïde et de l'ovocyte (Inoue *et al.*, 2005). Cette protéine est indispensable, il en existe néanmoins d'autres facilitant l'interaction entre le spermatozoïde et l'oocyte (Evans, 2002). Cette interaction gamétique, combinée à l'action de IZUMO1 et de son récepteur, permet la fusion de la membrane interne de l'acrosome avec la membrane plasmique de l'oocyte. A la suite de cette fusion, le noyau et certains organelles du spermatozoïde entrent dans l'oocyte, le noyau maternel complète sa seconde division de méiose, c'est le début du développement embryonnaire.

I.II.III.II : Les altérations fonctionnelles du spermatozoïde

A la lumière des différents éléments exposés ci-dessus, on peut remarquer qu'il y a un grand nombre d'acteurs indispensables au bon déroulement de la spermatogénèse, de la maturation du

spermatozoïde et de la fécondation. Il existe des altérations affectant la production ou la qualité des spermatozoïdes et qui peuvent être mises en évidence en réalisant un spermogramme (analyse microscopique du sperme et des spermatozoïdes).

(i) Les azoospermies sont caractérisées par l'absence totale de spermatozoïdes lors de l'examen microscopique et sont considérées comme les altérations les plus graves. On en discerne deux types, les Azoospermies Obstructives (OA), qui sont caractérisées par des défauts anatomiques de la sphère génitale sans pour autant avoir d'altérations de la spermatogénèse (Wosnitzer and Goldstein, 2014), et les Azoospermies Non Obstructives (NOA), qui elles sont caractérisées par une sphère génitale mâle physiologiquement normale associée à une incapacité à produire des spermatozoïdes. Les causes sont diverses : hormonales, infectieuses, génétiques.

(ii) Les oligozoospermies sont caractérisées par une faible concentration en spermaotozoïdes, pouvant avoir pour causes des anomalies du tractus génitale mâle, hormonales, génétiques et épigénétiques (McLachlan, 2013).

(iii) Les asthénozoospermies sont associées à la diminution ou à l'absence de motilité des spermatozoïdes.

(iv) Les tératozoospermies sont associées à la présence de nombreux spermatozoïdes présentant des anomalies morphologiques.

Les patients infertiles ont le plus souvent des défauts de spermogramme de type OAT (Oligo-Asthéno-Tératozoospermie) (Jungwirth *et al.*, 2012).Dans d'autres cas, il existe des altérations de la fertilité sans altérations du spermogramme. Il s'agit par exemple de l'absence de capacitation, d'absence d'hyperactivation, d'absence de réaction acrosomique, de dommages à l'ADN et d'absence d'interaction et de fusion avec l'ovocyte (Agarwal and Allamaneni, 2004; Pizzol *et al.,* 2014). Des tests complémentaires peuvent néanmoins être effectués en clinique, pour identifier des infertilités ayant ces causes.

Dans leur ensemble ces observations montrent qu'il existe un nombre important de causes d'infertilité affectant la physiologie du spermatozoïde, faisant des paramètres spermatiques de bons candidats pour prédire la fertilité des taureaux.

I.II.III.III : Les paramètres spermatiques ne permettent pas d'identifier tous les taureaux subfertiles

Dans les centres de collecte de semence, chaque éjaculat est inspecté à l'aide d'un microscope ou d'un CASA (Computer Assisted Semen Analysis) rapidement après l'éjaculation des taureaux. Cela permet d'analyser la présence ou non de spermatozoïdes, leur morphologie et leur motilité. De plus, étant donné que ces spermatozoïdes sont cryoconservés, certains contrôles qualité de la semence visent à évaluer les paramètres spermatiques après congélation et décongélation afin d'apprécier leur résistance à cette étape traumatisante. Ainsi, il est possible d'identifier des taureaux ayant des subfertilités ou infertilités provenant d'anomalies fonctionnelles de la semence. Cela démontre l'utilité de ces analyses pour identifier certaines formes d'infertilités ou de subfertilités.

Cependant, comme cela a été exposé plus haut, tous les troubles affectant la fonction du spermatozoïde ne se manifestent pas nécessairement lors d'un examen microscopique et il peut donc rester des taureaux subfertiles qui sont normozoospermiques (n'ayant aucune anomalie détectée lors du spermogramme), mais qui portent néanmoins des altérations affectant la fonctionnalité du spermatozoïde. Ainsi, parmi les taureaux retenus par les centres, il subsiste des individus de fertilité médiocre qui n'ont pas d'anomalies majeures au niveau de leur spermogramme.

C'est pour cela que des études ont tenté de prédire la fertilité sur la base de paramètres spermatiques non analysés lors des contrôles routiniers classiques, comme par exemple : la résistance membranaire, le potentiel mitochondrial, la fragmentation de l'ADN, la viabilité et la réaction acrosomique. L'étude de Holden et collègues n'a pas pu mettre en évidence de différences significatives entre taureaux fertiles et subfertiles sur l'intégralité de ces paramètres (Holden *et al.*, 2017). Une autre étude, où le statut oxydatif des spermatozoïdes a été analysé en plus des autres cités ci-dessus, confirme ces

résultats (Sellem *et al.*, 2015). Cette étude montre que pris individuellement chaque paramètre spermatique est faiblement corrélé à la fertilité des animaux. Cependant, cette étude a également montré qu'une combinaison optimale de ces paramètres permettait d'expliquer la variance phénotypique de la fertilité des animaux à hauteur de 39%, ce qui reste insuffisant.

Conclusion

Dans leur ensemble ces résultats montrent que premièrement, les paramètres spermatiques ne sont pas assez fiables pour prédire avec une précision suffisante la fertilité des taureaux. Deuxièmement, bien que la génétique soit source de candidats intéressants, il n'existe actuellement pas d'évaluation génomique de la fertilité mâle pratiquée en routine. De plus au vu de la faible héritabilité du TNR, on peut s'attendre à ce que la précision de la prédiction soit également faible à modérée. Ainsi, dans le but d'améliorer la qualité de prédiction, il pourrait être intéressant de s'intéresser à d'autres types de données biologiques liées à la fertilité. C'est le cas notamment des marques et mécanismes épigénétiques, dont le rôle dans la spermatogénèse, la physiologie du spermatozoïde et le développement embryonnaire est bien établi, en particulier chez les espèces modèles. Ainsi, prendre en compte cette information dans un modèle de prédiction pourrait potentiellement permettre d'expliquer une partie supplémentaire de la variance phénotypique de la fertilité mâle. La prochaine partie de cette introduction aura donc pour but de présenter différentes marques et mécanismes épigénétiques impliqués dans la fertilité et d'expliquer leur rôle dans la mise en place de ce phénotype.

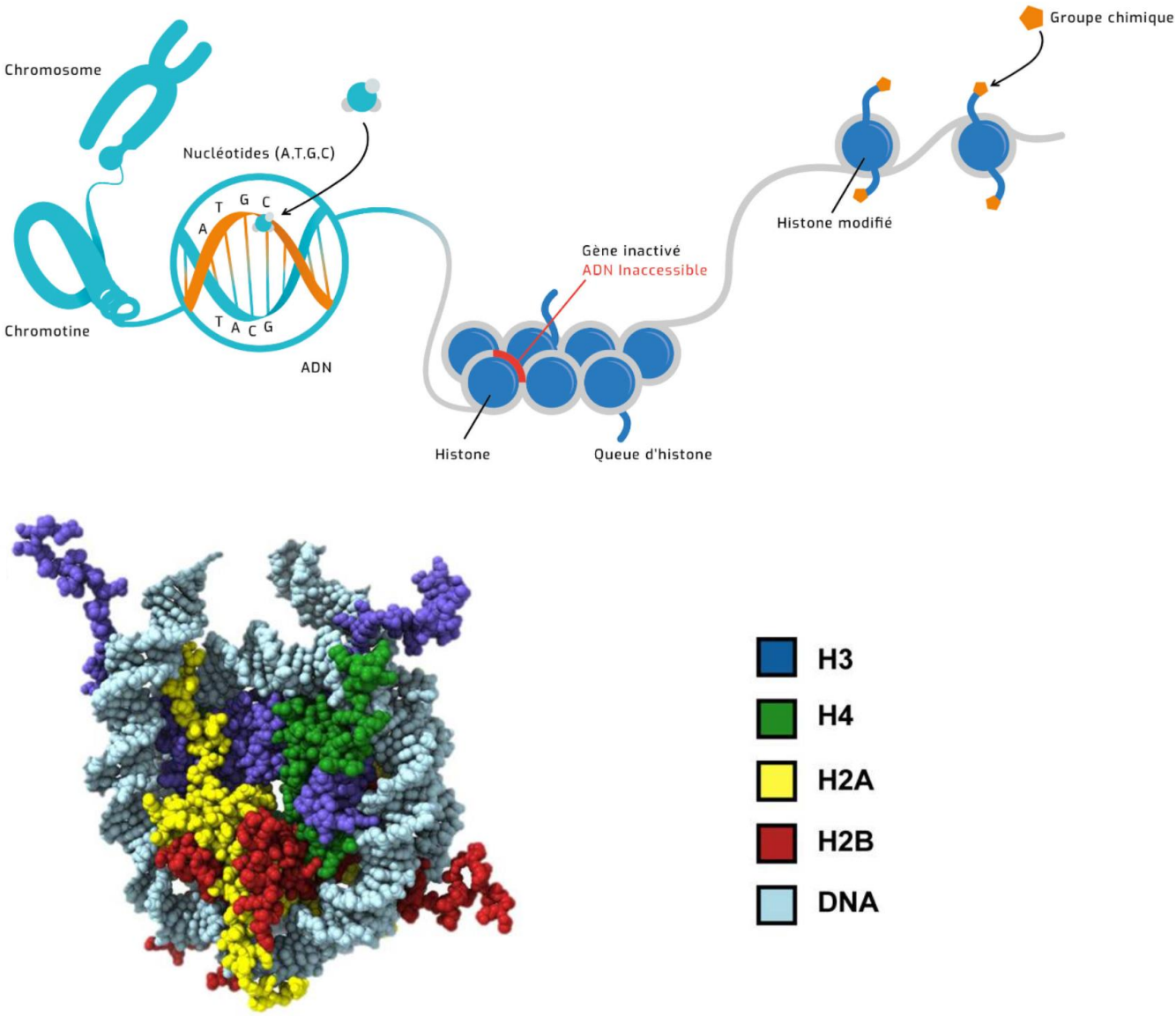


Figure 6 : La chromatine. Les chromosomes sont constitués de chromatine, qui est un assemblage des sous-unités des histones et de l'ADN. En fonction de modifications chimiques touchant l'ADN lui mêmes ou les sous unités d'histones, la chromatine peut se trouver dans un état compact (hétérochromatine) ce qui réduit ou empêche l'activité transcriptionnelle, ou relâché (euchromatine) favorisant la transcription.(Adapté de (Zhou, Gaullier, and Luger 2019))

II Epigénétique et fertilité mâle

L'épigénétique peut se définir par l'ensemble des marques apposées sur le génome, qui vont impacter la physiologie cellulaire sans pour autant modifier la séquence génétique de la cellule (Beaujean *et al.*, 2020). Ces marques sont héritables *a minima* entre deux générations cellulaires. Ce concept a dans un premier temps été proposé par Waddington dans le but d'illustrer les interactions complexes entre génotype et phénotype qui interviennent lors de la différenciation cellulaire et l'ontogénèse (Waddington, 2012). L'épigénétique permet en effet d'expliquer le fait qu'un organisme vivant métazoaire peut avoir différents types cellulaires, ayant des fonctions pouvant parfois être très différentes, tout en ayant le même génome. L'épigénétique est médiée par des acteurs comme les modifications post-traductionnelles des histones, la méthylation de l'ADN et les petits ARN non codants (sncRNA). Ces marques épigénétiques interagissent avec l'ADN ou ses produits, permettant de conférer une identité cellulaire héritable à travers la mitose, en définissant des états chromatiniens qui permettent de réguler l'expression des gènes et d'organiser l'architecture nucléaire en fonction du type cellulaire.

Cette partie sera consacrée à une description des modifications post-traductionnelles des histones, de la méthylation de l'ADN et des scnRNA, à leur rôle dans la différenciation des cellules germinales mâles et le développement embryonnaire, et aux variations de ces marques qui ont pu être décrites en lien avec la fertilité mâle. Comme une grande partie de ce travail a porté sur la méthylation de l'ADN, cette marque épigénétique a été décrite plus en détail.

II.I : Les histones

II.I.I : Les marques post-traductionnelles des histones

Dans les cellules eucaryotes l'ADN est enroulé autour de protéines : ce sont les histones (Kornberg and Lorch, 1999). Ces protéines sont de nature basique, ainsi l'ADN, de par ses propriétés physico-chimiques, possède naturellement une affinité pour les histones (Figure 6). L'interaction entre ces deux

éléments forme la chromatine. Dans les noyaux, la chromatine est caractérisée par des états différents : l'euchromatine, qui est une forme décondensée de la chromatine, riche en gènes et transcriptionnellement active ; à l'inverse l'hétérochromatine est une forme compacte de la chromatine. L'hétérochromatine est subdivisée en deux catégories : l'hétérochromatine constitutive qui est composée principalement de régions répétées et pauvres en gènes ; et l'hétérochromatine facultative, pouvant elle, basculer dynamiquement vers un état actif et riche en gènes non exprimés.

Ces états chromatiniens permettent la régulation de l'expression des gènes, la maintenance du génome et sa structure spatiale. Ils sont définis en partie par des modifications chimiques des histones. Dans les cellules somatiques, il existe 5 types d'histones différents : H1, H2A, H2B, H3 et H4 (Phillips and Johns, 1965). Les unités H2A, H2B, H3, H4 sont chacune présente en double exemplaire et forment un octamère autour duquel 142 paires de bases d'ADN s'enroulent. L'histone H1 permet de stabiliser la structure globale. L'ensemble de ce complexe s'appelle le nucléosome.

Les unités d'histones H2A, H2B, H3 et H4 présentent des extrémités qui dépassent de la surface des nucléosomes, ce qui les rend accessibles à différentes enzymes capables de modifier leurs propriétés physico-chimiques (Luger and Richmond, 1998). Il existe une très grande diversité dans les modifications post-traductionnelles d'histones. En effet, chaque sous unité d'histone possédant une extrémité peut être modifiée, ces modifications peuvent avoir lieu sur plusieurs types d'acides aminés et peuvent être de nature chimique différente (méthylation, acétylation etc). Il existe donc une grande diversité possible dans le codage de l'information par les marques post traductionnelles d'histones. Les différents types de modifications et leurs conséquences sur la chromatine et l'expression des gènes sont détaillés dans la revue de (Sadakierska-Chudy and Filip, 2015). A la lumière de la diversité de ces marques, une nomenclature explicite existe pour comprendre directement la nature de la modification qui impacte l'histone d'intérêt. Par exemple, la marque H3K9me3 signifie qu'il s'agit d'une modification de l'histone H3, sur la neuvième lysine de la queue d'histone et qu'elle consiste en l'ajout de trois groupements méthyle. Cette marque est en général associée à de l'hétérochromatine

constitutive et est souvent associée à la méthylation de l'ADN, en particulier au niveau des régions péri-centromériques et centromériques (Moore *et al.,* 2013).

II.I.II : L'implication des marques post-traductionnelles d'histones dans la fertilité

Cependant, le spermatozoïde est une cellule à la chromatine atypique, et la formation de la chromatine spermatique est un processus hautement régulé (Blanco and Cocquet, 2019). Le spermatozoïde est presque dépourvu d'histones, et la proportion d'histones restantes est espèce dépendante. En effet, au cours de la spermiogénèse, une très grande partie des histones est remplacée par des protamines, petites protéines très basiques et riches en arginines et en cystéines (Bao and Bedford, 2016). Comme exposé dans la première partie de l'introduction, ces protéines permettent une condensation plus importante de l'ADN spermatique en formant des structures d'ordre supérieur, ce qui le protège du stress oxydant de l'épididyme par exemple. Les protamines établissent des ponts disulfure inter- et intramoléculaires lors du transit à travers l'épididyme, qui consolident cette structure nucléaire (Noblanc *et al.,* 2014). Malgré le remplacement d'une grande partie des histones dans le spermatozoïde, les marques d'histones jouent un rôle important dans la formation des gamètes mâles au cours de la spermatogénèse.

En effet, au cours de la spermatogénèse, les histones présentes dans les spermatogonies, les spermatocytes et les spermatides rondes n'ont pas été encore remplacées par les protamines et peuvent donc être modifiées afin de moduler l'organisation chromatinienne. Il a en effet été montré qu'au cours de ce processus, les modifications post-traductionnelles sont très contrôlées. C'est le cas notamment pour les modifications affectant H3 et H4 (Godmann *et al.*, 2007). Il a par exemple été montré qu'en inhibant SUV39H, une méthyltransferase permettant la méthylation en H3K9, une diminution de la formation d'hétérochromatine constitutive est observée, qui a pour conséquences d'augmenter le taux d'erreurs lors des méioses et de réduire la fertilité (Peters *et al.*, 2001). De plus au moment de l'élongation des spermatides, une augmentation de l'acétylation des histones se produit

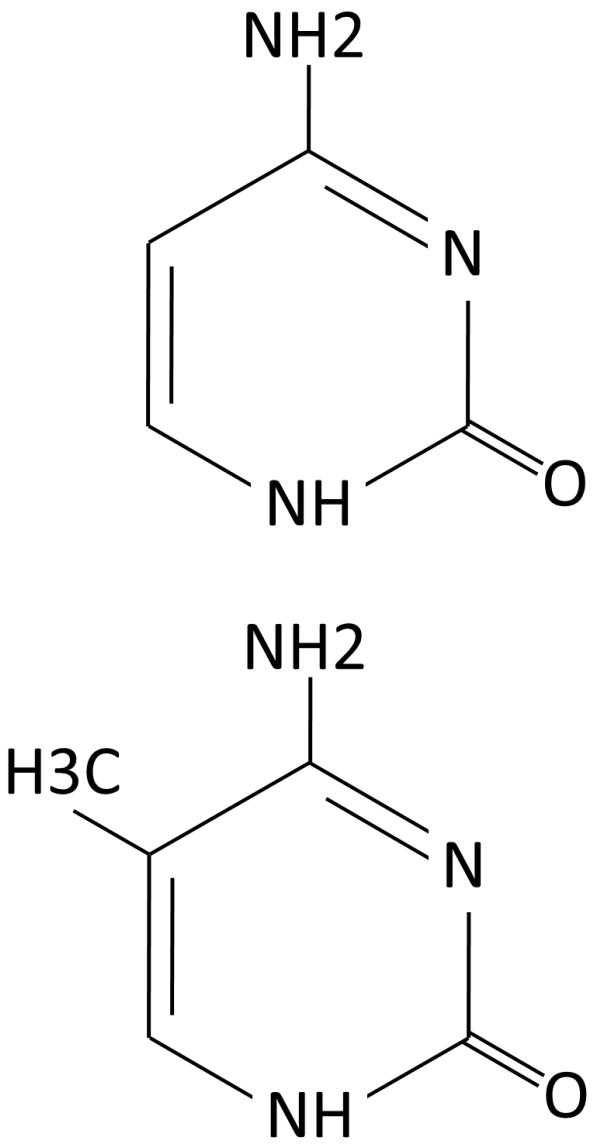


Figure 7 : La cytosine et la 5’ méthyl-cytosine. Schéma représentant une cytosine (en haut) et une méthyl-cytosine (en bas).

afin de conférer un statut « relâché » à la chromatine, permettant de faciliter le remplacement des histones par les protamines (Hazzouri *et al.*, 2000). Enfin, bien que pendant la spermiogénèse une grande partie des histones soit remplacée par des protamines, il existe certaines régions contenant des histones résiduelles portant des modifications post traductionnelles, qui pourraient jouer un rôle lors du développement embryonnaire (Gatewood *et al.*, 1987; Pittoggi *et al.*, 1999; Hammoud *et al.*, 2009, 2011).

II.II : La méthylation de l'ADN

II.II.I : Machinerie enzymatique en lien avec la méthylation de l'ADN

La méthylation de l'ADN est une marque épigénétique décrite pour la première fois en 1948 (Hotchkiss, 1948). Sa particularité est qu'elle modifie directement l'ADN en apposant de manière covalente un groupement méthyle sur le cinquième carbone des cytosines (Figure 7). Cette modification s'effectue la plupart du temps dans des contextes génétiques qualifiés de CpG –c'est-à-dire une cytosine directement suivie par une guanine. Il existe aussi des méthylations hors contexte CpG qui ne seront pas abordées dans ce chapitre.

Comme toutes les marques épigénétiques la méthylation de l'ADN est de nature dynamique, en particulier lors de la différenciation cellulaire mais également en fonction des conditions environnementales et physiologiques. Par exemple, l'âge affecte les capacités de l'organisme à maintenir de manière fidèle la méthylation de l'ADN dans un type cellulaire donné (Horvath, 2013). Elle est cependant moins dynamique que les modifications post-traductionnelles des histones présentées dans le chapitre précédent. Les protéines permettant l'apposition des groupements méthyles sur les cytosines font partie de la famille des DNMT pour « DNA methyltransferases ». Chez les bovins comme chez de nombreux mammifères il en existe 4 différentes : DNMT1, DNMT3A, DNMT3B et DNM3L (DNMT3C étant spécifique des rongeurs) (Lyko, 2018).

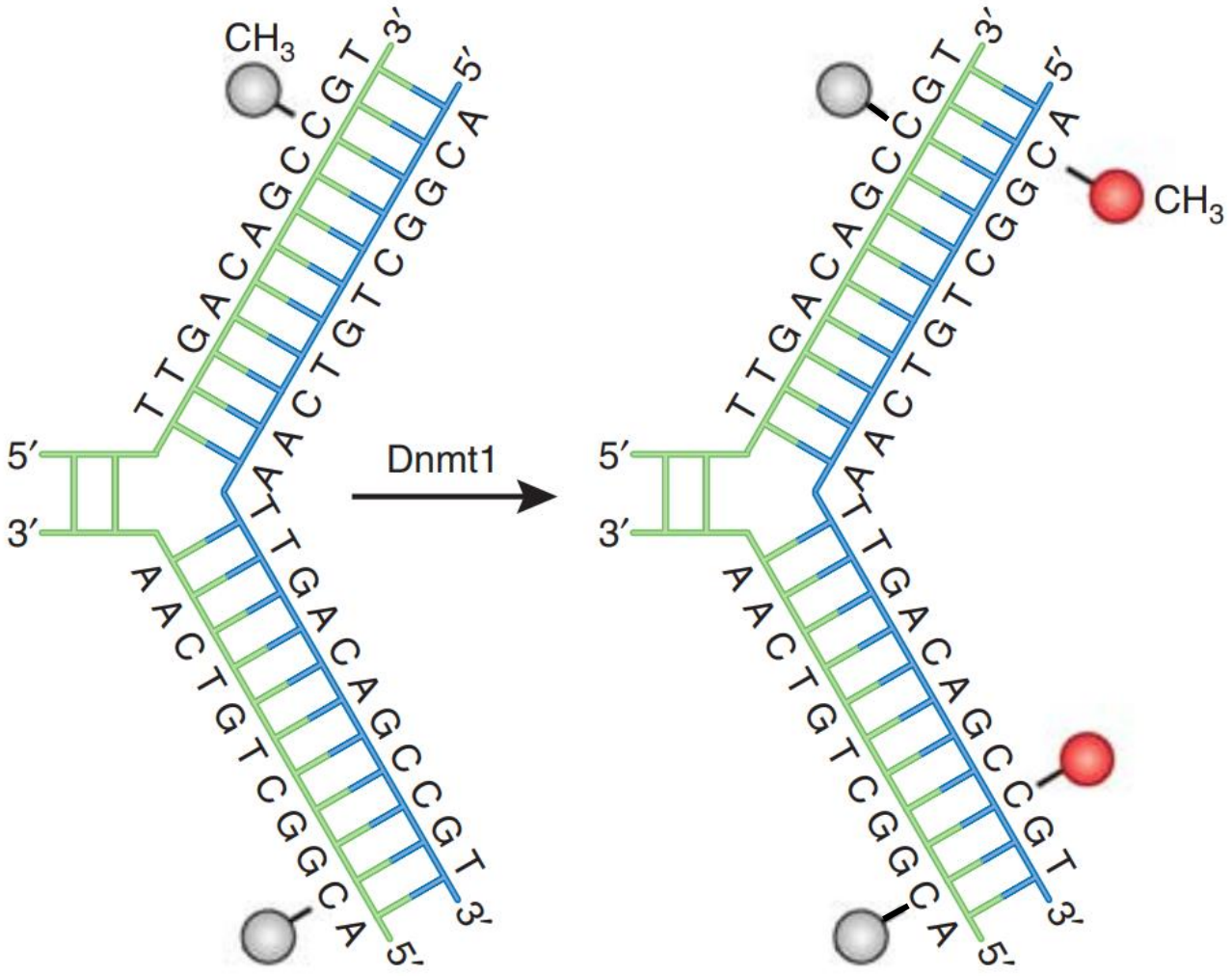


Figure 8 : Transmission de la méthylation de l'ADN par l'enzyme DNMT1. Schéma illustrant le mécanisme de maintenance de la méthylation de l'ADN. Au cours de la division cellulaire, au moment de la réplication de l'ADN, le brin d'ADN issu de la cellule mère (en vert) sert de matrice pour la synthèse du brin néo-synthétisé (en bleu). Si une cytosine méthylée est présente sur le brin provenant de la cellule mère, alors UHRF1 de par son affinité avec les sites hémi-méthylés va se fixer sur l'ADN est entrainer le recrutement de la DNMT1. De ce fait la méthylation de l'ADN va pouvoir être recopiée sur le brin néo-synthétisé est ainsi être présente dans les deux cellules filles après la division cellulaire (adapté de (Moore, Le, and Fan 2013)).

DNMT1 est la méthyltransferase de maintenance. Comme expliqué précédemment, une caractéristique importante de l'épigénétique est son aspect héritable entre chaque mitose. Le rôle de DNMT1 est de permettre aux cellules filles d'hériter du méthylome (profil de méthylation de l'ADN) de la cellule mère. Au moment de la réplication de l'ADN précédant la division cellulaire, elle va se localiser au niveau de la fourche de réplication grâce à son interaction avec le co-facteur UHRF1, là où réside le brin d'ADN d'origine (Figure 8) et le brin d'ADN néo synthétisé qui est vierge de toute méthylation sur les cytosines (Leonhardt *et al.*, 1992). Les marques de méthylation présentes sur les cytosines du brin d'ADN mère vont alors être recopiées par la DNMT1 sur le brin néosynthétisé (Pradhan *et al.*, 1999; Hermann *et al.,* 2004). Ainsi, les deux cellules filles héritent du profil de méthylation présent originellement dans la cellule mère. En plus de son activité de maintenance au cours de la division cellulaire, cette protéine permet également de restaurer la méthylation d'un site qui aurait été affecté par des lésions à l'ADN (Mortusewicz *et al.*, 2005).

DNMT1 est donc une enzyme permettant la maintenance du signal de méthylation de l'ADN. En amont de cette maintenance se situe la méthylation qualifiée de *de novo*, c'est-à-dire intervenant sur un site CpG dénué de méthylation sur les 2 brins complémentaires. Cette action est réalisée par un couple d'enzymes que sont la DNMT3A et la DNMT3B (Okano *et al.*, 1999).

La dernière protéine impliquée dans la méthylation *de novo* de l'ADN est la protéine DNMT3L. Cette protéine a la particularité de ne pas posséder de site catalytique et ne peut donc pas apposer de groupement méthyle sur les cytosines. Cependant elle est quand même appelée methyltransférase car elle s'associe à la DNMT3A et DNMT3B pour guider la méthylation *de novo* (Hata *et al.*, 2002). Son expression est prépondérante dans les cellules germinales, dans lesquelles la méthylation *de novo* de l'ADN est particulièrement intense.

La nature dynamique de la méthylation de l'ADN implique à la fois des phénomènes d'apposition et d'effacement. Il existe deux mécanismes permettant à terme d'éliminer le groupement méthyle d'une cytosine afin de perdre localement le signal de méthylation de l'ADN. Ces deux mécanismes diffèrent

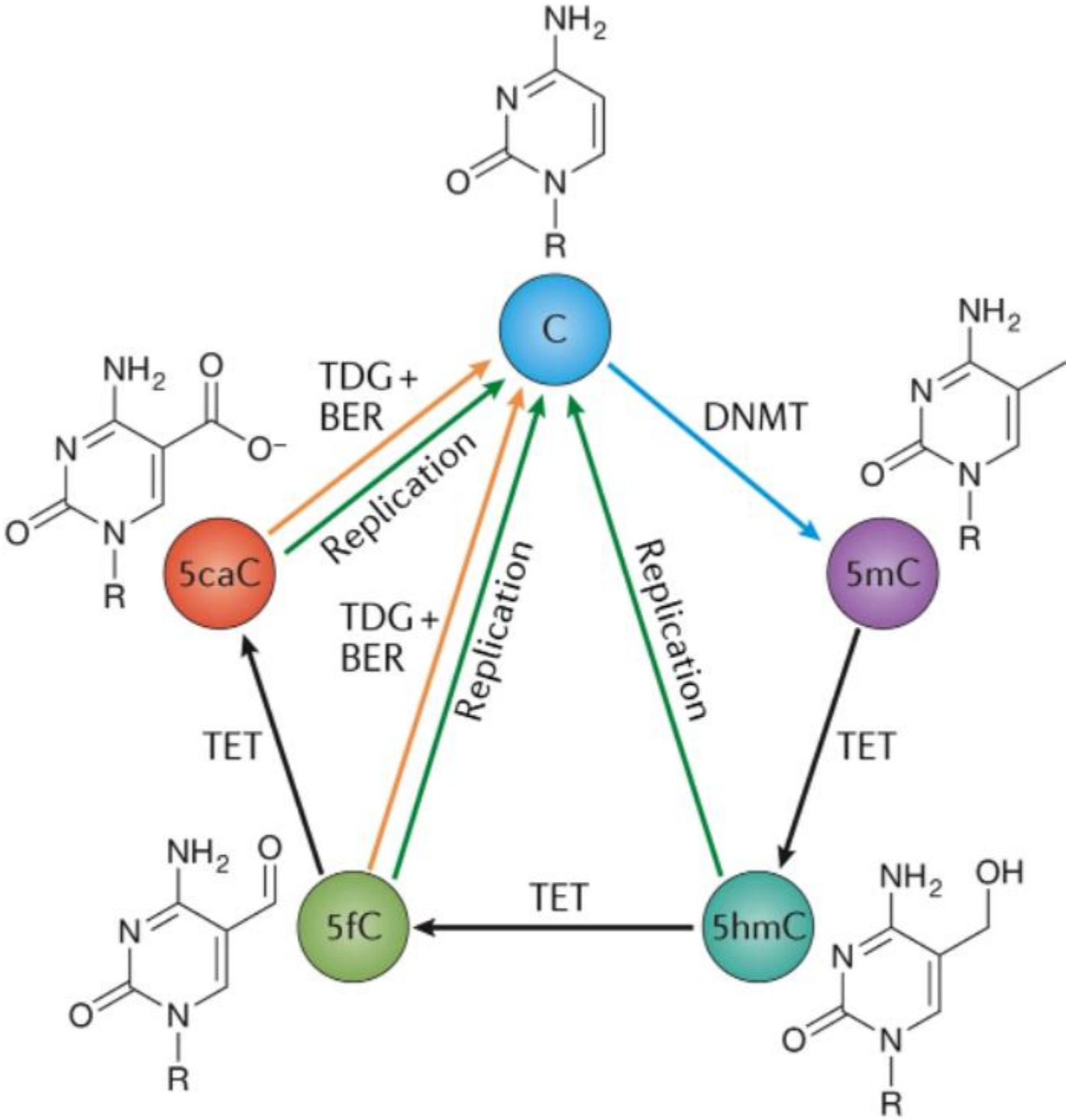


Figure 9 : Schéma illustrant la voie active de déméthylation médiée par les enzymes TET. Les cytosines peuvent être méthylées par les protéines de la famille des DNMT et former des méthylcytosines (flèche bleu). Ce groupement méthyle peut être supprimé par un mécanisme actif impliquant les enzymes de la famille des TET. Leur action permet de modifier les méthylcytosines en hydroxymethylcytosine (5hmC), formylcytosine (5fC) et carboxylcytosine (5caC) (flèche noir). Les 5fC et les 5caC peuvent être remplacées par une cytosine non modifiée par le mécanisme impliquant les BER (« Base Excision Repair ») (flèche orange). Les 5hmC, 5fC et 5caC, n'étant pas reconnues par l'enzyme de maintenance DNMT1, elles peuvent être diluées passivement au fil des divisions mitotiques (flèche verte). Inspiré de (Wu and Zhang 2017)

significativement, puisque l'un est passif tandis que l'autre est actif. Le mécanisme passif s'applique lorsque la DNMT1 est inactive. Ainsi, lors de la réplication de l'ADN au moment de la division cellulaire, le profil de méthylation n'est pas copié sur le nouveau brin d'ADN. Cela a pour conséquence de produire deux cellules filles mosaïques en termes de méthylation. Au fur et à mesure des divisions, le nombre de cellules présentant des méthylations originaires de la cellule mère diminue, jusqu'à ce que le niveau de méthylation soit complétement minoritaire, voire effacé. Ce mécanisme passif est particulièrement efficace en phase d'intense activité mitotique, ce qui est le cas du développement embryonnaire. Le mécanisme actif est médié par des enzymes permettant de supprimer la liaison covalente entre le cinquième carbone de la cytosine et le groupement méthyle. Ces enzymes font parties de la famille des TET (« Ten Eleven Transclocation methylcytosine dyoxygenase protein »), et les différentes étapes de la déméthylation active sont présentées Figure 9 (Wu and Zhang, 2017).

La méthylation de l'ADN permet également le recrutement de protéines ayant des fonctions dans la régulation transcriptionnelle, qui sont qualifiées de « readers ». Cette propriété est directement liée à l'action de la méthylation de l'ADN dans la répression transcriptionnelle qui est détaillée dans le chapitre suivant. Il existe trois familles de « readers » associées à la méthylation de l'ADN : les protéines de la famille MBD (« Methyl CpG Binding Domain »), les UHRF et certaines protéines avec un motif en doigt de zinc (Kaiso, ZBTB4, ZBTB8). Les MBD et les protéines à doigt de zinc sont recrutées sur l'ADN par des CpG méthylés, et participent à la répression transcriptionnelle en recrutant des corépresseurs transcriptionnels et des protéines modifiant les modifications post traductionnelles des histones afin de promouvoir la formation d'hétérochromatine (Sasai *et al.,* 2010; Du *et al.*, 2015). Les protéines de la famille des UHRF interagissent avec les sites hémi-méthylés, et ont une fonction de co-facteurs de la DNMT1 de façon à répliquer le signal de méthylation au moment de la mitose (Bashtrykov *et al.*, 2014).

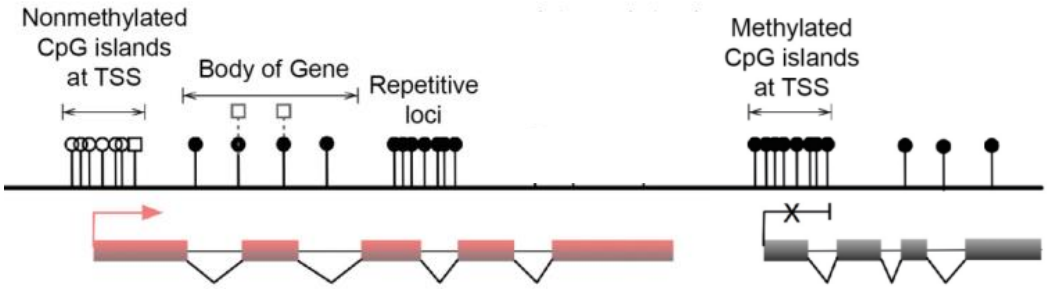


Figure 10 : Régulation de la transcription par la méthylation de l'ADN. Schéma représentant les modes d'action de la méthylation de l'ADN sur la transcription. De gauche à droite : Lorsque les îlots CpG présents dans les régions promotrices des gènes sont fortement déméthylés, et que ces îlots sont associés à d'autres marques épigénétiques ou facteurs de transcription, cela permet la transcription du gène proximal. Les CpG méthylés dans le corps de gènes hors îlots CpG favorisent l'activité transcriptionnelle des gènes. Les séquences répétées transposables dans le génome sont des régions riches en îlots CpG, souvent hautement méthylées pour les maintenir dans un état de répression transcriptionnelle. Lorsque les îlots CpG présents dans les régions promotrices des gènes sont hyperméthylés, cela a pour conséquence de réprimer l'activité transcriptionnelle.

II.II.II : Fonctions biologiques de la méthylation de l'ADN

La méthylation de l'ADN est souvent analysée pour son rôle dans la régulation de l'expression des gènes codants, cependant ce n'est pas son seul rôle. La section suivante a pour but d'exposer les différents rôles biologiques de la méthylation de l'ADN dans les cellules de mammifères.

II.II.II.I : Régulation de la transcription

Comme exposé précédemment, la méthylation de l'ADN est présente de façon majoritaire au niveau des sites CpG. Les sites CpG sont sous-représentés dans les génomes comparés à leur occurrence théorique. Une explication associée à ce fait est la propriété des cytosines méthylées à être 14 fois plus mutagènes que les autres nucléotides (Schübeler, 2015). Malgré cela, il existe des régions où la concentration des sites CpG est très importante comparé à une répartition théorique aléatoire : ce sont les îlots CpG.

Environ 70% des gènes possèdent des îlots CpG dans leurs région promotrice (Saxonov *et al.,* 2006). Presque paradoxalement, les gènes ayant des séquences promotrices localisées dans des îlots CpG sont le plus souvent hypométhylés au niveau de ces îlots dans les différents types cellulaires et quel que soit leur niveau d'expression, suggérant donc une régulation par d'autres facteurs. Une des hypothèses possibles pour expliquer cette hypométhylation est la pression de sélection qui pourrait s'exercer pour que ces sites CpG soient non méthylés, et donc conservés au cours de l'évolution (Angeloni and Bogdanovic, 2021). Malgré cela, quand les îlots CpG associés à des prometteurs sont hyperméthylées, ils entraînent une répression transcriptionnelle du gène associé (Figure 10). Ce phénomène est décrit dans des contextes précis comme par exemple le verrouillage des gènes de pluripotence lors de la différenciation cellulaire, l'inactivation du chromosome X, ou le maintien de l'empreinte parentale (détaillé plus bas). Dans un contexte pathologique, c'est aussi le cas de la répression des gènes suppresseurs de tumeurs qui peut être à l'origine de cancers.

L'action de la méthylation de l'ADN est associée à la compaction de la chromatine sous forme d'hétérochromatine facultative. Cette compaction empêche la machinerie transcriptionnelle d'accéder à l'ADN et ainsi promeut l'inhibition transcriptionnelle dans les régions promotrices (Brenet *et al.*, 2011). La méthylation de l'ADN n'est cependant pas essentielle à la répression transcriptionnelle, car il existe des gènes présentant des îlots CpG non méthylés au niveau de leur prometteur et qui restent inactifs. De plus pour une action efficace sur la transcription, la méthylation de l'ADN doit plutôt être localisée dans des régions denses en CpG (par exemple, à proximité des îlots CpG, dans les « shores » et les « shelves »), car des promoteurs de gènes localisés dans des régions peu denses en CpG mais méthylées sont plutôt transcriptionnellement actifs (Weber *et al.*, 2007). A l'inverse, la méthylation de l'ADN dans le corps des gènes (hors îlots CpG) est plutôt associée à une activation transcriptionnelle (Aran *et al.*, 2011). Elle pourrait être impliquée dans la répression de sites illégitimes de la transcription, et dans le ralentissement de la progression de l'ARN polymérase (Neri *et al.*, 2017).

II.II.II.II : Répression des éléments répétés

La régulation transcriptionnelle par la méthylation de l'ADN ne se limite pas aux gènes, mais est également importante pour la régulation des éléments répétés transposables et structuraux.

Les éléments répétés transposables représentent presque la moitié du génome chez les mammifères. Les éléments transposables ont la particularité d'être mobiles au sein du génome et sont connus pour être des vecteurs de l'évolution (Kazazian, 2004). De plus, d'un point de vue physiologique, ces éléments peuvent également avoir des fonctions importantes lorsqu'ils sont transcrits dans un cadre très particulier, c'est le cas par exemple des transposons de type « long interspersed nuclear elements » (LINE) qui permettent de réguler l'accessibilité de la chromatine chez la souris après la fécondation (Jachowicz *et al.*, 2017) et semblent également finement régulés lors de la spermatogenèse (Blythe *et al.*, 2021). Cependant, de par leur capacité à s'insérer dans l'ADN, ces éléments peuvent être dangereux s'ils ne sont pas maintenus à l'état inactif, avec des conséquences allant de l'induction de mutations à l'envahissement du génome. En effet, si le site d'insertion est

contenu dans un fragment d'ADN important, cela peut potentiellement annihiler sa fonction et le rendre néfaste pour la survie de la cellule et de l'organisme (Kuster *et al.*, 1997). Ainsi, les éléments répétés transposables doivent être finement régulés, et la méthylation de l'ADN est un acteur de cette régulation (Yoder *et al.,* 1997).

Enfin certaines séquences répétées, comme les séquences satellites, ont des rôles structuraux importants et participent à la formation et au maintien des centromères et des télomères. La méthylation de l'ADN intervient dans la stabilité des centromères (Scelfo and Fachinetti, 2019). En effet, elle joue un rôle dans la protection des centromères lors des recombinaisons homologues qui ont lieu au cours de la méiose, ainsi qu'en assurant le positionnement correct des kinétochores lors des mitoses afin d'empêcher les aneuploïdies. De plus, l'hypométhylation des régions péri-centromériques est associée à la pathologie ICF (« Centromeric Instability and Facial anomalies »). Cette pathologie rare est dans de nombreux cas causée par une mutation dans le gène de la DNMT3B. Enfin, la méthylation de l'ADN joue un rôle dans la stabilité des régions télomériques (revue dans (Toubiana and Selig, 2020)). Cette régulation se fait par la méthylation des régions subtélomériques, ce qui permet de réguler un long ARN non codant (lncRNA) : TERRA. Dans les cas d'ICF, qui sont caractérisés également par une hypométhylation des régions subtélomériques, on observe une augmentation des recombinaisons homologues et une diminution de la taille des télomères entre générations cellulaires et par conséquent une accélération de la sénescence (vieillissement cellulaire).

II.II.II.III : Inactivation du chromosome X

La méthylation de l'ADN est également impliquée dans le mécanisme appelé inactivation du chromosome X chez les mammifères. Ce mécanisme a pour but de compenser la présence du chromosome X supplémentaire dans les cellules de sexe femelle (génotype XX) par rapport aux cellules de sexe mâle (génotype XY). En effet, le chromosome X est un grand chromosome portant un grand nombre de gènes, ce qui n'est pas le cas du chromosome Y. Le fait qu'il soit porté en deux exemplaires chez les femelles et en un seul exemplaire chez les mâles crée donc un déséquilibre que l'inactivation

d'un des deux chromosomes X permet de compenser chez les femelles. Sans cette inactivation il y aurait une transcription beaucoup plus importante des gènes de ce chromosome chez les femelles par rapport aux mâles. Cela engendrerait de graves conséquences sur le développement embryonnaire (Borensztein *et al.*, 2017). Ce mécanisme se met en place très tôt au moment du développement embryonnaire. Il est initié par la transcription d'un long ARN non codant : *Xist* (Brown *et al.*, 1991, 1992). Cet ARN non codant va recouvrir un des deux chromosomes X et va permettre son inactivation transcriptionnelle, ne laissant qu'un seul chromosome X actif dans les cellules (Clemson *et al.*, 1996; Chaumeil *et al.*, 2006). Après le signal déclencheur de l'inactivation du chromosome, des mécanismes vont rentrer en jeu afin de maintenir ce signal au sein d'une cellule et au cours des divisions cellulaires. C'est à ce moment que la méthylation de l'ADN va intervenir. Les ilots CpG du chromosome X sont hautement méthylés afin de garantir sa répression transcriptionnelle (Norris, Brockdorff and Rastan, 1991). Il a de plus été montré que des altérations de la méthylation de l'ADN du chromosome X inactivé permettaient une réactivation transcriptionnelle de ce dernier, soulignant donc l'importance de sa régulation par la méthylation de l'ADN (Csankovszki *et al.,* 2001).

II.II.II.IV : Empreinte parentale

Enfin, la méthylation de l'ADN est également impliquée dans l'empreinte parentale. Ce mécanisme biologique correspond au fait qu'un seul allèle s'exprime, parmi les deux présents dans une cellule pour un gène donné. La copie du gène s'exprimant est dépendante de son origine parentale selon un mécanisme qui n'est pas aléatoire. En effet, pour un gène donné le même allèle est toujours exprimé (celui provenant de la mère ou du père). Ce phénomène a dans un premier temps été mis en évidence en générant des embryons gynogénotes et androgénotes, c'est-à-dire ayant respectivement deux copies de chromosomes maternels ou paternels, et en observant une létalité embryonnaire (McGrath and Solter, 1984). Le premier gène soumis à empreinte découvert a été *IGF2* (Barlow *et al.*, 1991). Ce gène a une empreinte maternelle, c'est-à-dire qu'il est uniquement exprimé par le chromosome paternel et est réprimé sur le chromosome maternel. La mise en place de l'empreinte parentale est

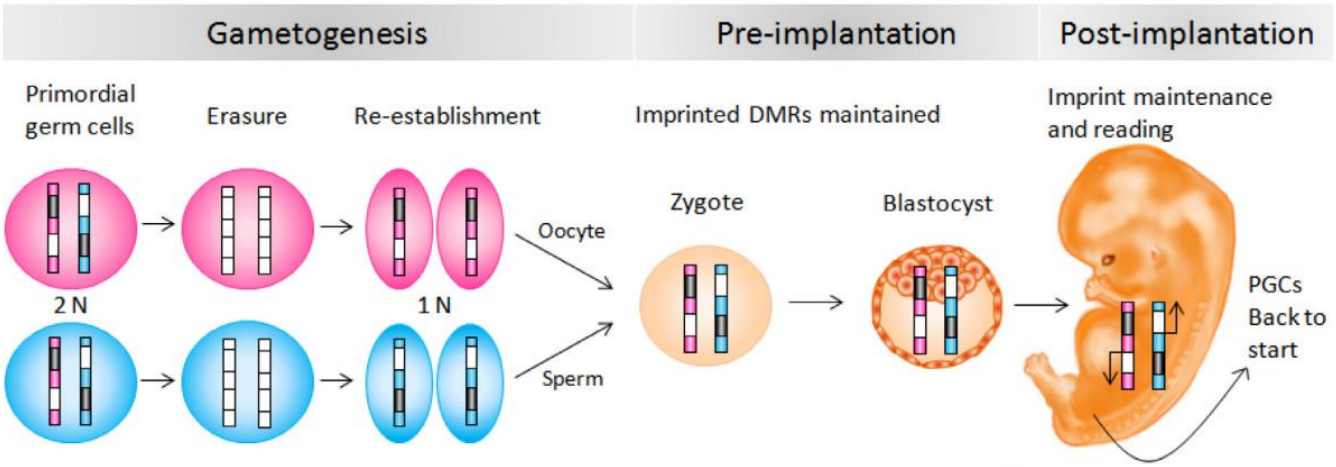


Figure 11 : Mise en place de l'empreinte parentale. Après la fécondation, et après quelques étapes de différenciation cellulaire, certaines cellules de l'épiblaste vont migrer au niveau des crêtes génitales et former les « Primordial Germ Cell » (précurseurs de la lignée germinale). Les gènes soumis à empreinte parentale (dont l'expression monoallélique est dépendante de l'origine parentale) se trouvent différentiellement méthylés en fonction de l'origine parentale de l'allèle. Rapidement après la mise en place du lignage PGC, une vague de déméthylation touche ces cellules, ce qui a pour conséquence l'effacement sexe spécifique des ICR. Cet effacement permet d'effacer la marque parentale de ces allèles. Au cours de la vague de méthylation *de novo* touchant aussi bien la voie femelle que la voie mâle, une ces marques sont réapposées en fonction du sexe de l'individu en formation. Ces marques seront maintenues jusqu'à la production des ovocytes et des spermatozoïdes. Ainsi au moment de la fécondation, le zygote aura un allèle paternel et un allèle maternel portant chacun le patron de méthylation adéquat (inspiré de (Ishida and Moore 2013))

assurée par plusieurs marques épigénétiques, dont la méthylation de l'ADN (Kelsey and Feil, 2013). (Kelsey and Feil 2013). Au cours de la différenciation cellulaire des cellules germinales mâles, la méthylation de l'ADN est apposée sur les deux allèles des gènes soumis à empreinte paternelle, au niveau de « centres d'empreintes » A l'inverse, les « centres d'empreintes » des gènes à empreinte maternelle resteront dépourvus de méthylation. Cela permettra d'assurer une expression physiologique de ces gènes afin de garantir des conditions normales pour le développement (Figure 11). En effet, beaucoup de gènes soumis à empreinte connus (une centaine chez l'Homme, moins chez le bovin) sont impliqués dans la régulation de la croissance fœtale.

II.II.III Reprogrammation de la méthylation de l'ADN : de la différenciation des cellules germinales mâles aux premières étapes du développement

Au cours de la partie précédente, nous avons pu apprécier les rôles généraux de la méthylation de l'ADN chez les mammifères. La partie suivante aura pour but d'exposer le rôle de la méthylation de l'ADN directement en lien avec la fertilité mâle. Mais tout d'abord, la dynamique de la méthylation de l'ADN au cours de la spermatogénèse puis après la fécondation est abordée ci-dessous.

La fertilité, c'est-à-dire la capacité des individus à produire une descendance, repose sur trois processus : la différenciation des gamètes mâles et femelles, la fécondation et le développement jusqu'à la naissance. Après l'implantation de l'embryon dans la muqueuse utérine, le développement repose en grande partie sur les capacités maternelles à mener une gestation. C'est pourquoi ce paragraphe s'intéressera essentiellement aux étapes précoces du développement.

Il se trouve justement que le méthylome est drastiquement reconfiguré pendant la différenciation des cellules germinales et le développement précoce. Ainsi toute altération de la méthylation de l'ADN à ces stades peut conduire à des défauts de la gamétogénèse ou de l'embryogenèse.

Ce chapitre a donc pour objectif d'exposer la dynamique de la méthylation de l'ADN, qui subit deux reprogrammations (effacement et réapposition), du lignage cellulaire originel des cellules germinales

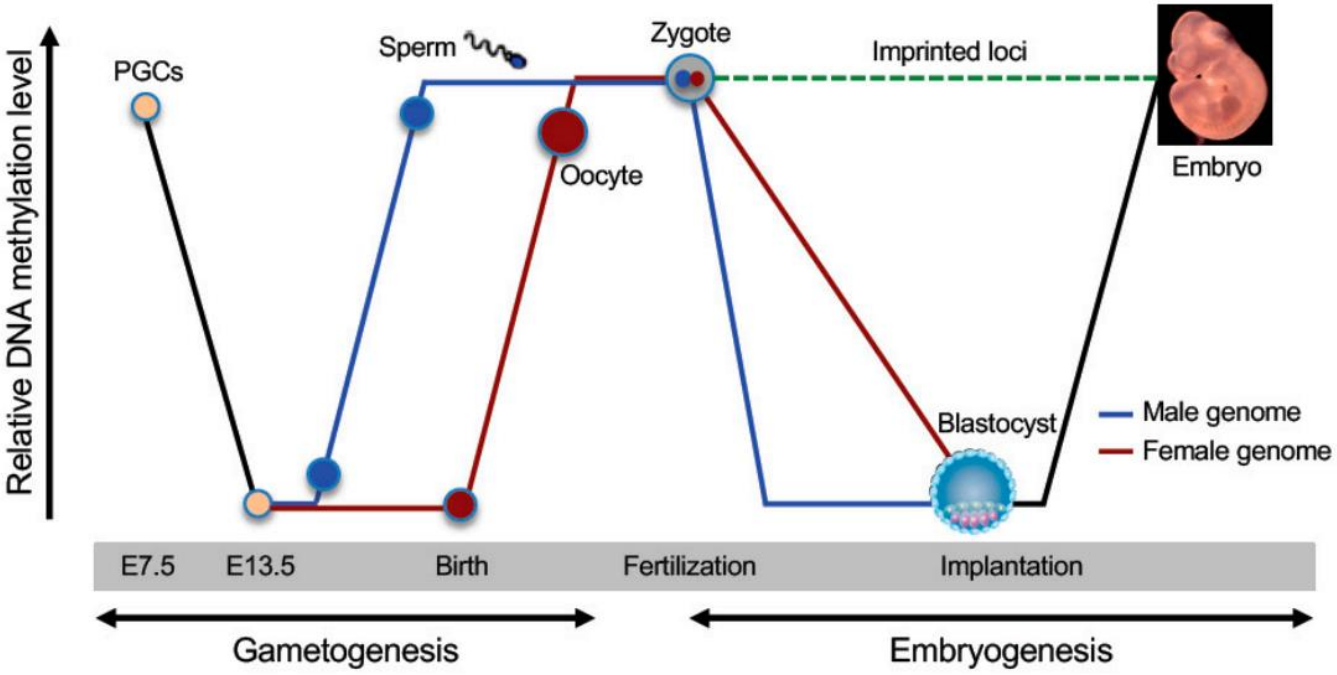


Figure 12 : Les deux vagues de reprogrammation au cours du développement. La première vague a lieu après la formation du lignage des PGC aux alentours de 8 jpc. Une déméthylation pan-génomique permet d'effacer l'épigénome somatique acquis au cours de la différenciation de l'épiblaste ainsi que l'empreinte au niveau des ICR. Une méthylation *de novo* fera suite à cette déméthylation et sa dynamique dépendra du sexe de l'individu en formation. Chez la souris mâle, la méthylation *de novo* commence à 13.5 jpc et une grande partie de ce méthylome sera acquis pendant le développement fœtal. Rapidement après la fécondation, les génomes maternels et paternels vont à nouveau être reprogrammés avec une dynamique de déméthylation différentes. Cette fois-ci les régions soumises à empreinte ne sont pas reprogrammées. Chez la souris, le minimum de méthylation est atteint au stade blastocyste. Une méthylation *de novo* est ensuite réapposée en fonction des différents lignages cellulaires. (Zeng and Chen 2019)

mâles jusqu'aux premiers stades du développement embryonnaire (Figure 12). Les études sur lesquelles s'appuient les connaissances actuelles ont principalement été réalisées chez la souris, ainsi, sauf mention contraire, les temps indiqués dans cette partie se réfèrent à cette espèce. A noter que chez la souris, le temps de gestation est de 20 jours seulement. L'unité utilisée pour décrire les temps de gestation est le *jour post coitum* (jpc).

Période fœtale :

Les cellules donnant naissance aux gamètes au cours du développement sont les cellules germinales primordiales (« primordial germ cells », PGC). Ce lignage cellulaire se met en place à partir de l'épiblaste 6,25 jours après la fécondation chez la souris (Chiquoine, 1954), puis migre au niveau des crêtes génitales de l'embryon (précurseurs des gonades), jusqu'à 13,5 jours après la fécondation. A ce stade, les crêtes génitales sont colonisées par les PGC, qui arrêtent progressivement de se diviser et commencent à se différencier en pro-spermatogonies. Au cours de la migration des PGC et de la colonisation des crêtes génitales, c'est-à-dire entre 8 et 13,5 jours après la fécondation, des changements drastiques affectent la méthylation de l'ADN, qui est presque intégralement effacée pour atteindre un minimum de 4 % à 13,5 jours (elle était de 74 % dans les cellules progénitrices des PGC à 6,5 jpc) (Seisenberger *et al.*, 2012; Kobayashi *et al.*, 2013). Cette déméthylation est nécessaire pour configurer un état permettant aux PGC d'une part d'exprimer les gènes de différenciation germinale et d'autre part de supprimer l'empreinte parentale héritée après la fécondation, afin de pouvoir rétablir un nouveau marquage parental correspondant au sexe de l'individu en formation sur les 2 allèles (les cellules germinales mâles étant à ce stade encore diploïdes). Cette déméthylation s'effectue en deux phases ; la première est marquée par une déméthylation passive. En effet, BLIMP1 et PRDM14, deux protéines permettant la spécification de l'épiblaste en PGC, vont inhiber UHRF1, DNMT3A et DNMT3B (Kagiwada *et al.*, 2013; Magnúsdóttir *et al.*, 2013). Cette inhibition de la machinerie de maintenance et de méthylation *de novo* permet ainsi une dilution du signal de méthylation au cours des divisions cellulaires. La déméthyltion passive est ensuite directement suivie par une deuxième

vague de déméthylation active médiée par les enzymes tet1 et tet2 (Hackett *et al.*, 2013). Malgré cette déméthylation, une faible proportion du génome reste néanmoins méthylée. Il s'agit principalement de régions répétées transposables, montrant encore une fois l'importance de la méthylation de l'ADN dans la maintenance du génome (Lees-Murdock *et al.,* 2003).

A l'issue de cette déméthylation, une vague de reméthylation des prospermatogonies se produit à partir de 13,5 jpc. Cette vague de méthylation *de novo* est principalement médiée par DNMT3A et DNMT3L dont l'expression est importante dans les prospermatogonies après 13,5 jpc, contrairement à DNMT1 et DNMT3B qui sont faiblement exprimées à ces stades (La Salle *et al.*, 2004). Pour que la reméthylation soit efficacement dirigée vers les éléments transposables et éviter ainsi l'invasion du génome déméthylé, ces DNMT coopèrent avec des petits ARN non codants, les piRNA dont la fonction sera détaillée plus tard dans le manuscrit, présentant des homologies de séquence avec les transposons (Aravin *et al.*, 2008; Kuramochi-Miyagawa *et al.*, 2008). Des problèmes au niveau de cette méthylation *de novo,* qui intervient essentiellement au stade fœtal, peuvent conduire à des blocages méiotiques après la puberté et rendre les animaux stériles, comme c'est le cas lors de l'inactivation des gènes DNMT3L (Bourc'his and Bestor, 2004) et DNMT3A (Kaneda *et al.*, 2004) dans la lignée germinale mâle. C'est au cours de cette vague de reméthylation que l'empreinte parentale mâle se met en place (Li *et al.*, 2004).

<u>Période post-natale :</u>

Aux alentours de 3 jours après la naissance des souris, les prospermatogonies reprennent une activité mitotique (elles étaient bloquées en phase G1 de mitose depuis 16,5 jpc), afin de former les spermatogonies. Cette reprise des cycles cellulaires est accompagnée d'une augmentation de la transcription des enzymes DNMT1 et DNMT3B afin de maintenir le statut de méthylation *de novo* acquis précédemment (La Salle *et al.*, 2004).

Une grande partie de la reméthylation des PGC se déroule en période fœtale car en trois jours, de 13.5 jpc à 16.5 jpc, le taux de méthylation augmente de 4 à 50% (Seisenberger *et al.*, 2012). Néanmoins une

étude récente s'est intéressée à la dynamique de méthylation chez le rat aux stades PGC, prospermatogonies, spermatogonies à P10 (10 jours après la naissance), spermatocytes, spermatides rondes et spermatozoïdes présents dans la tête ou la queue de l'épididyme (Ben Maamar *et al.*, 2022). Des analyses différentielles de méthylation ont été réalisées entre les 2 stades consécutifs et le plus grand nombre de DMR (Régions Différentiellement Méthylées) a été identifié au moment du passage des prospermatogonies en spermatogonies. Ces résultats montrent que d'importantes modifications de méthylation ont également lieu dans les cellules germinales après la naissance. De plus, au stade spermatogonie, le méthylome des cellules germinales n'est pas figé, car des changements sont encore détectés au cours de la spermatogénèse (Oakes *et al.*, 2007; Ben Maamar *et al.*, 2022). En plus de cette évolution au cours de la spermatogénèse, il a été mis en évidence que le méthylome spermatique évoluait lors du transit par l'épididyme et également tout au long de la vie de l'individu (Lambert *et al.*, 2018; Takeda *et al.*, 2019; Chen *et al.*, 2022).

Comparés aux cellules somatiques, les spermatozoïdes éjaculés sont des cellules hypométhylées chez diverses espèces, et les gènes hypométhylés jouent un rôle important dans la maturation des cellules germinales (Molaro *et al.*, 2011). Cette méthylation plus faible touche en particulier les séquences satellites, dont le statut est particulièrement hypométhylé dans les spermatozoïdes de taureaux (Perrier *et al.*, 2018).

<u>Après la fécondation :</u>

A la suite de la fécondation, le zygote possède une copie du génome paternel et une copie du génome maternel avec leurs épigénomes respectifs. Rapidement après cette fécondation, les deux génomes subissent une nouvelle vague de déméthylation, mais avec une dynamique et une amplitude différentes pour le génome maternel et pour le génome paternel (Mayer *et al.*, 2000). Cette vague de déméthylation a pour but d'effacer la méthylation de l'ADN gamétique, afin de permettre au zygote de retrouver un état de totipotence, qui pourra donner vie aux différents lignages cellulaires lors du développement embryonnaire. Alors que la déméthylation du génome maternel est essentiellement

passive et suit le rythme des clivages cellulaires, la déméthylation du génome d'origine paternelle est réalisée à la fois de manière active et passive. En effet au stade zygote la disparition de la méthylation de l'ADN paternel est rapide et corrélée à l'apparition de l'hydroxyméthylation (Inoue and Zhang, 2011), un phénomène médié par l'enzyme TET3 (Iqbal *et al.*, 2011). Une fois la 5mC modifiée en 5hmC, 5fC et 5CaC (voir Figure 9), une excision active et une dilution passive de ces marques au cours des divisions de l'embryogénèse pourraient entrainer la déméthylation (Inoue and Zhang, 2011; Inoue *et al.*, 2011; Kohli and Zhang, 2013). A l'instar de la vague de déméthylation touchant les PGC, il existe également un échappement de certaines régions génomiques à cette vague de reprogrammation épigénétique. Encore une fois une grande partie des régions échappant à cette reprogrammation concerne des éléments transposables, car leur insertion dans le génome au stade unicellulaire pourrait compromettre le développement physiologique de l'embryon (Lane *et al.*, 2003). Cependant une différence notable par rapport à la reprogrammation des PGC est le fait que l'empreinte parentale paternelle et maternelle doit être maintenue. Des mécanismes spécifiques interviennent pour permettre aux marques de méthylation permettant cette empreinte de résister à la reprogrammation. Ces mécanismes pourraient également protéger d'autres éléments du génome, puisque l'ampleur de la déméthylation est moindre que lors de la reprogrammation des PGC.

Le niveau le plus bas de méthylation de l'ADN est atteint au stade de blastocyste à 3,5 jpc chez la souris (Wang *et al.*, 2014) et au stade 6 à 8 cellules chez le bovins (Dobbs *et al.*, 2013; Jiang *et al.*, 2018). Rapidement après, une méthylation *de novo* de l'ADN est médié par les enzymes DNMT3A et DNMT3B (Okano *et al.*, 1999; Auclair *et al.*, 2014; Wang *et al.*, 2014). A 6,5 jpc le niveau de méthylation de l'ADN dans l'épiblaste (partie de l'embryon qui reste pluripotente) a atteint son maximum. Ce profil de méthylation évoluera néanmoins différemment dans les différents types cellulaires, afin de leur configurer et de maintenir une identité tout au long du développement et au cours de la vie de l'individu.

Ces différents éléments montrent l'importance de la méthylation de l'ADN à la fois dans le lignage cellulaire dont les gamètes sont issus, permettant une fécondation, et également dans le développement embryonnaire lui-même.

II.II.IV : Implication de la méthylation de l'ADN dans la fertilité mâle

Comme exposé précédemment, l'inactivation de gènes de la machinerie de méthylation dans les cellules germinales mâles entraine une stérilité, démontrant clairement que la méthylation de l'ADN est une modification épigénétique indispensable pour la fertilité mâle. Cependant ces résultats ont été obtenus chez la souris dans des conditions expérimentales. Cette partie a donc pour but d'illustrer les différents problèmes de fertilité rencontrés chez l'homme en lien avec une altération du méthylome spermatique. Ensuite, l'état de l'art concernant des modifications du méthylome spermatique en lien avec des différentiels de fertilité chez le taureau sera abordé.

Chez l'homme, une revue exhaustive de la littérature sur l'infertilité et/ou la subfertilité en association avec la méthylation de l'ADN a été publiée en 2020 par Fredrika Asenius et collègues (Åsenius *et al.,* 2020). Les études mentionnées dans cette revue ont en général pour objet des groupes de patients contrôles (qualifiés de fertiles) contre des patients ayant des problèmes de fertilité avec ou sans altération de la spermatogénèse (reflétée par la mesure des paramètres spermatiques). Cette partie du manuscrit aura donc pour but de synthétiser les principaux travaux couverts par cette revue ainsi que d'autres études non citées par cette dernière ou plus récentes.

La première étude a comparé des spermatozoïdes ayant une bonne mobilité et ceux ayant une mauvaise mobilité chez des donneurs fertiles (Barzideh *et al.,* 2013). La principale conclusion est que les spermatozoïdes ayant une mauvaise mobilité ont un taux de méthylation supérieur comparé aux spermatozoïdes ayant une bonne mobilité. Cela suggère qu'une hyperméthylation est associée à la production de spermatozoïdes de mauvaise qualité.

Cette étude a été réalisée à un niveau de méthylation globale et ne permet donc pas de mettre en évidence les éléments génomiques impliqués dans ce différentiel de méthylation. D'autres études ont été réalisées à une résolution plus précise, et mettent en évidence un différentiel de méthylation à l'échelle du gène ou du CpG. Les régions soumises à empreinte parentale ont ainsi été abondamment étudiées dans le contexte de la fertilité masculine. L'un des gènes à empreinte les plus étudiés est *H19* qui est à empreinte paternelle. Le centre d'empreinte de *H19* est significativement plus faiblement méthylé pour des patients ayant des anomalies de spermogramme comparés à des individus normozoospermiques (Marques *et al.*, 2004, 2008; Dong *et al.*, 2017). Ces travaux sont en adéquation avec la méta-analyse réalisée par D. Santi et collègues en 2017, où 18 études portant sur l'analyse de *H19* avaient mis en lumière une hypométhylation spermatique de ce gène dans des infertilités masculines (Santi *et al.*, 2017).

Un autre gène soumis à empreinte ayant reçu de l'attention est le gène *MEST*, qui a également une méthylation paternelle. Trois études ont mis en évidence une hypométhylation de ce gène chez des patients présentant une oligozoospermie (Marques *et al.*, 2008; Poplinski *et al.*, 2010; Kläver *et al.*, 2013). Il existe cependant une étude contradictoire ne montrant pas de lien entre problème de fertilité et le statut de méthylation de *MEST* (Marques *et al.*, 2004). Néanmoins, la méta analyse conduite par Santi et collègues a mis en évidence une hyperméthylation significative chez les patients infertiles (Santi *et al.*, 2017).

Enfin *SNRPN*, un gène soumis à empreinte maternelle, a également été mis en évidence dans la méta analyse de Santi avec une hyperméthylation chez les patients infertiles (Santi *et al.*, 2017)

Les gènes soumis à empreinte sont certes importants dans l'analyse des défauts de méthylation de l'ADN pouvant causer une infertilité masculine, cependant certains autres gènes ont également des fonctions essentielles dans la spermatogénèse et le développement embryonnaire. Ainsi de nombreuses études se sont intéressées à cette question chez l'homme avec une approche gène candidat, c'est-à-dire en regardant le statut de méthylation d'un seul gène à la fois en fonction de la

fertilité. Un des gènes non soumis à empreinte le plus étudié avec une approche gène candidat est *MTHFR* –un gène impliqué dans le métabolisme du folate-, où l'hyperméthylation du promoteur est associée à des problèmes de spermatogénèse (Wu *et al.*, 2010; Tian *et al.*, 2014; Rezaeian *et al.,* 2021).

Cependant les approches gène candidat présentent la faiblesse de n'interroger qu'un ou quelques gènes à la fois avec un *a priori* important. On peut tout à fait imaginer que l'infertilité n'est pas la conséquence d'un problème de méthylation sur un seul gène mais sur plusieurs. De plus, les régions analysées sont la plupart du temps bien caractérisées, ce qui exclut *de facto* l'analyse de nouvelles régions pouvant être importantes pour la fertilité mâle. C'est pour cela, et grâce à l'apparition d'approches pan-génomiques, que d'autres études ont interrogé un grand nombre de CpG en association avec des infertilités, souvent à l'aide de puces de méthylation. Jusqu'à présent aucune signature d'infertilité ou de subfertilité commune à ces études n'a été mise en évidence (Pacheco *et al.*, 2011; Schütte *et al.*, 2013; Camprubí *et al.*, 2016; Jenkins *et al.*, 2016; Laqqan *et al.,* 2017[a,b;c;d],2018). Une des causes possibles de cette absence de signature commune peut être la diversité phénotypique existant dans les infertilités analysées (défauts de motilité, teratozoospermies, infertilité avec un profil normozoospermique). Néanmoins ces études ont mis en évidence des gènes différentiellement méthylés présentant un lien avec la spermatogénèse ou le développement embryonnaire. Cela peut suggérer une forme d'hétérogénéité au niveau des gènes impactés, mais une homogénéité dans les fonctions des gènes mis en évidence.

II.II.V : Implication de la méthylation de l'ADN dans la subfertilité bovine

Comme expliqué dans le chapitre I, la fertilité bovine est un sujet majeur de préoccupation dans les élevages et les filières professionnelles, en partie du fait de son importance économique. Devant l'absence d'évaluation génomique de la fertilité mâle en routine et l'imprécision des estimations de la fertilité des taureaux basées sur leurs paramètres spermatiques, le méthylome spermatique des taureaux a été envisagé comme une nouvelle source de biomarqueurs de la fertilité mâle bovine.

Actuellement 6 analyses pan-génomiques de la méthylation ont été publiées dans ce but, dont les principaux résultats sont décrits ci-dessous.

La première étude a été publiée chez le buffle en 2014 par Verma et collègues (Verma *et al.*, 2014). Une comparaison du méthylome spermatique de 5 taureaux fertiles contre 5 taureaux subfertiles a été réalisée. Après validation du différentiel de fertilité par FIV (Fécondation *in vitro*), les auteurs ont pu identifier un total de 96 gènes différentiellement méthylés ayant une fonction dans la différenciation des cellules germinales, la spermatogénèse, la capacitation et le développement embryonnaire. Une étude de Gross et collègues en 2020, analysant le méthylome spermatique de taureaux contrastés en terme de fertilité, a également permis l'identification de gènes différentiellement méthylés en fonction de la fertilité mâle (Gross *et al.,* 2020). Enfin, dans une étude parue en 2021, Takeda et collègues ont analysé des taureaux de fertilité contrastée de race Japanese Black sur une puce humaine dont peu de positions sont exploitables chez le bovin. Cette étude a mis en évidence 143 DMC (Cytosines Différentiellement Méthylées) et également montré qu'il était possible d'établir une relation linéaire entre le niveau de méthylation de certaines de ces DMC et la fertilité des animaux (Takeda *et al.*, 2021).

L'étude de Kropp et collègues parue en 2017 avait également pour but de comparer le méthylome spermatique de taureaux ayant une fertilité contrastée, mais aussi d'utiliser la semence de ces taureaux en FIV afin de suivre le développement des embryons jusqu'au stade blastocyste et d'analyser leur transcriptome (Kropp *et al.*, 2017). L'étude n'a pas mis en évidence de différences significatives de taux de clivage et de taux de formation de blastocystes entre les taureaux fertiles et subfertiles. Cependant, malgré cette absence de différences en termes de qualité d'embryons sur les critères énoncés ci-dessus, 98 gènes étaient différentiellement exprimés entre embryons générés à partir des taureaux fertiles ou subfertiles. Certains de ces gènes sont impliqués dans le développement embryonnaire au-delà du stade blastocyste, suggérant que le suivi sur une plus longue période permettrait potentiellement de mettre en évidence des différences de développement entre

descendance des taureaux fertiles et subfertiles. En plus de ces analyses transcriptomiques, un profilage du méthylome spermatique a mis en évidence 76 régions différentiellement méthylées entre groupes de fertilité, touchant des gènes impliqués dans la spermatogénèse et le développement embryonnaire.

En 2019, Fang et collègues ont analysés le méthylome spermatique de 3 taureaux fertiles et 7 taureaux subfertiles, classifiés à partir du SCR (Fang *et al.*, 2019). Ils ont dans un premier temps identifié des PMD (« Partially Methylated Domain »), qui sont des régions de plusieurs kilobases, en général peu méthylées et associées avec des marques post-traductionnelles d'histone représentatives de l'hétérochromatine. Après avoir réalisé une comparaison avec les PMD trouvés chez les animaux fertiles ou subfertiles, ils ont mis en évidence que ceux qui se trouvaient spécifiquement chez les animaux subfertiles étaient associés à des fonctions biologiques en lien avec la physiologie du spermatozoïde. Ils ont également réalisé une analyse différentielle de méthylation entres groupes de fertilités et ont mis évidence 11 663 DMR. Une partie de ces DMR co-localisent avec des régions génomiques ayant été mises en évidence par des analyses d'association génétique avec des caractères de fertilité.

Enfin, une étude récemment réalisée par Narud et collègues est parue en 2021 (Narud *et al.*, 2021). L'objectif de cette étude était la caractérisation du méthylome spermatique de taureaux contrastés en termes de fertilité, ainsi que l'appréciation de la fragmentation de l'ADN spermatique, et enfin la réalisation de FIV à l'aide des semences de taureaux des 2 cohortes. Ces recherches ont démontré que la fertilité des animaux était négativement corrélée à la qualité de l'ADN dans les spermatozoïdes, fournissant une explication potentielle pour au moins une partie de la variation de la fertilité des animaux. Ce résultat a été renforcé par le fait qu'il existe également une corrélation négative entre l'intégrité de l'ADN et les résultats de la FIV (taux de clivages et taux de blastocystes). En comparant le méthylome spermatique, les auteurs ont pu mettre en évidence 16 542 DMC entre catégories de

fertilité. Les gènes impactés par le différentiel de méthylation sont également impliqués dans la spermatogénèse et le développement embryonnaire.

II.III : Les petits ARN non codants

Les ARN non codants sont une classe de molécule dont l'appellation s'oppose directement au terme d'ARN « codants ». Les ARN codants contiennent une phase ouverte de lecture permettant la traduction en peptides ou protéines, et ce sont ces derniers qui ont une action biologique dans la cellule ou dans l'organisme. Par opposition, les ARN non codants sont également issus de la transcription de l'ADN mais ne sont pas traduits en protéines ; c'est alors la molécule d'ARN non codant qui est fonctionnelle. Il existe deux grandes catégories d'ARN non codants : les longs ARN non codants (lncRNA), ayant en général une taille de plus de 300 nucléotides, et les petits ARN non codants (sncRNA), ayant une taille comprise entre 25 et 40 nucléotides. Dans ce travail de thèse, et dans cette partie, nous nous intéresserons uniquement aux sncRNA, car pour des raisons méthodologiques il est difficile d'étudier les deux catégories à la fois. Il existe plusieurs familles de sncRNA, parmi lesquelles les miRNA (micro-ARN), les piRNA (ARN intéragissant avec PIWI), les tsRNA (fragments dérivés des ARN de transfert) et les rsRNA (fragments dérivés des ARN ribosomiques) seront décrits plus en détail. Cette partie a pour objectif de présenter de manière brève ces différents sncRNA présents dans les cellules germinales mâles et d'exposer leurs rôles dans la fertilité mâle.

II.III.I : Les miRNA

Les miRNA sont des sncRNA d'une longueur moyenne de 22 nucléotides (Ozsolak *et al.,* 2008). Les séquences d'ADN qui les codent sont souvent présentes dans les régions intragéniques, en particulier au niveau des introns (Ozsolak *et al.*, 2008). Les séquences des miRNA sont transcrites en pré-miRNA par l'ARN polymérase de type II, et sont généralement longues de quelques milliers de paires de bases avec une structure en tige boucle (Lee *et al.*, 2004; Ha and Kim, 2014). Après quelques étapes de maturation nucléaire, le pré-miRNA va être expulsé du noyau par l'exportine 5 afin d'être pris en charge par la protéine DICER qui va permettre son clivage afin que la taille du miRNA soit d'une

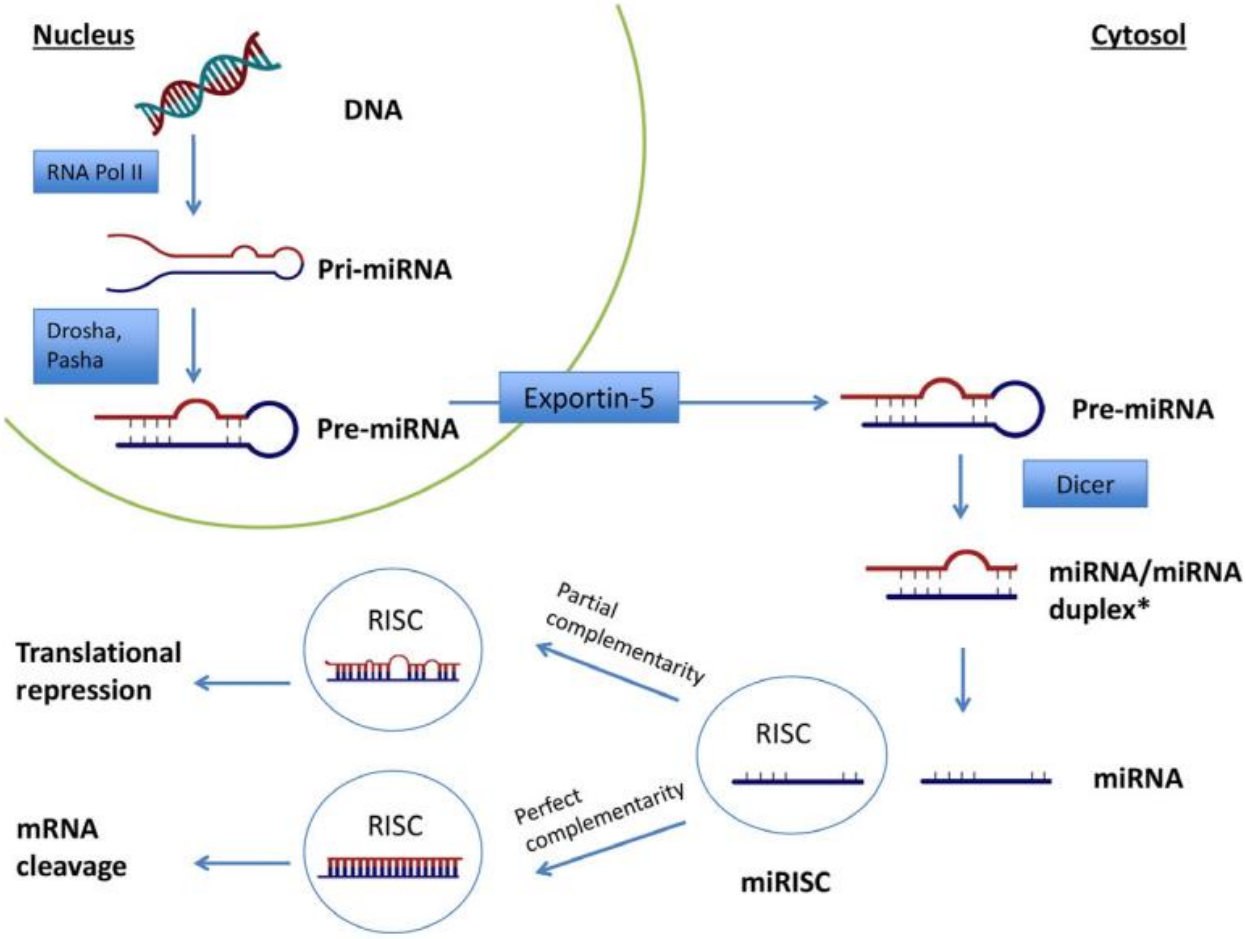


Figure 13 : Génération et mécanismes d'actions des miRNA. Les pré-miRNA sont transcrits par l'ARN polymérase de type II, maturés par Drosha et Posha est sont exportés du noyau vers le cytosplasme par l'exportine 5. Ils sont ensuite clivés par DICER et pris en charge par des protéines de la famille des Argonautes pour former le complexe RISC. Ce complexe va permettre de réaliser un contrôle post-transcriptionnelle des ARN, en bloquant la traduction ou clivant les ARNm. Inspiré de (Magri, Vanoli, and Corti 2018)

vingtaine de nucléotides, et que l'ARN soit sous forme mono-brin (Hutvágner *et al.*, 2001; Lund *et al.*, 2004). Le miRNA clivé et simple-brin interagit avec son ARN cible grâce à la complémentarité de séquence et avec des protéines Argonautes afin de former le complexe RISC – « RNA Induced Silencing Complex » - (Zealy *et al.*, 2017). Ce complexe a pour fonction d'inhiber l'expression des gènes, via la dégradation de leur ARN messager ou le blocage de la traduction. Après reconnaissance de l'ARN messager par le complexe RISC, celui-ci enclenche sa dégradation ou son blocage traductionnel (Figure 13). Ainsi, une diminution de la quantité d'ARNm et/ou de la synthèse protéique d'un gène est observée sans pour autant qu'il n'y ait d'impact sur sa transcription.

II.III.II : Les piRNA

Les piRNA, à l'instar des miRNA, ont également pour but de réguler la quantité de certains ARN dans les cellules. Les piRNA ont principalement une fonction dans l'inhibition des séquences transposables (Vagin *et al.*, 2006) mais également d'autres fonctions dans la formation de l'hétéchromatine, la régulation des ARN messagers, la régulation des lncRNA et la protection de l'intégrité du génome (Wang and Lin, 2021). La génération de piRNA se fait selon deux mécanismes. La première voie ressemble en partie au mécanisme mis en jeu pour la synthèse des miRNA : la séquence d'ADN codant pour le piRNA est transcrite, puis, après une étape de maturation nucléaire l'ARN est relâché dans le cytoplasme où il est clivé en plusieurs fragments par la protéine Zuc. Ces fragments sont ensuite chargés sur des protéines de la famille des Argonautes : les PIWI. Une fois cette interaction mise en place, l'ARN subit encore quelques modifications avant de former un piRNA mature (Saito *et al.*, 2007). Le second mécanisme s'appelle le mécanisme ping-pong, il permet de générer une réponse forte des piRNA lorsqu'un transposon est actif (Czech and Hannon, 2016). Cela est permis car, lorsqu'un transposon est identifié par un piRNA par le mécanisme décrit plus haut, le transposon est clivé mais n'est pas dégradé totalement. A la place, les fragments du transposon qui sont complémentaires du piRNA sont pris en charge par AGO3 (famille des Argonautes), pour générer de nouveaux piRNA spécifiques du transposon actif et ainsi amplifier la réponse.

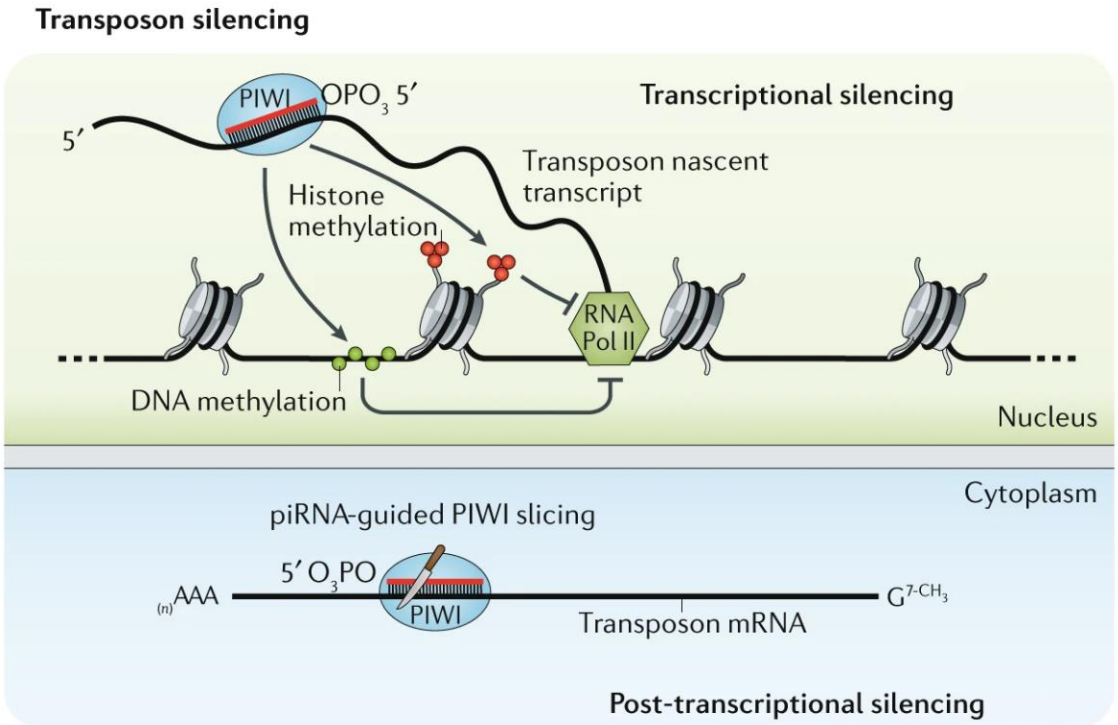


Figure 14 : Deux stratégies permettent aux piRNA de réguler la production de rétrotransposons dans les cellules. La première stratégie est la régulation transcriptionnelle. Le piRNA est présent dans le noyau, identifie le transposon en cours de transcription dont il est complémentaire et permet le recrutement de protéines promouvant la méthylation de l'ADN et des histones. Cela engendre une compaction de la chromatine afin de réprimer la transcription du transposon. La seconde stratégie fait appel à la régulation post transcriptionnelle. Certains piRNA reste dans le cytoplasme et détectent la présence de transposons transcrits afin de les cliver (Ozata et al. 2019)

Les piRNA régulent l'expression des transposons par deux mécanismes distincts, qui sont dictés par la nature de la protéine PIWI avec laquelle interagit le piRNA (Ozata *et al.*, 2019). Lors d'interactions avec le protéine MIWI2, le piRNA mature retourne dans le noyau afin de réprimer les transposons actifs en favorisant la méthylation des histones (H3K9) ainsi qu'en recrutant la machinerie de méthylation de l'ADN. La seconde action se déroule cette fois-ci non pas le noyau mais dans le cytoplasme, en clivant les transcrits des transposons actifs (Figure 14).

II.III.III : Les rsRNA et tsRNA

Les rsRNA et les tsRNA sont deux types d'ARN générés à partir d'autres ARN ayant des fonctions biologiques bien connues. Les tsRNA sont des fragments d'une longueur de 20 à 35 nucléotides issus de la dégradation des ARN de transferts –ARN chargés en acides aminés et reconnaissant un codon, ce qui permet la correspondance entre le signal nucléotidique et un signal peptidique lors de la traduction des ARNm. Les rsRNA sont des produits de clivage des ARN ribosomiques –ARN constituant le ribosome. Dans les deux cas, la génération de ces sncRNA est un mécanisme actif médié par différentes enzymes (Li *et al.,* 2018; Lambert *et al.,* 2019). Même si leur fonction n'est pas encore très bien caractérisée, plusieurs études montrent que ces deux types d'ARN interagissent avec la protéine Argonaute, suggérant un rôle dans la répression de l'expression des gènes via la dégradation d'ARN messager (Wei *et al.*, 2013; Kuscu *et al.*, 2018). Il a d'ailleurs été montré dans certaines études que la présence de ces sncRNA était directement liée à de la régulation de l'expression génique (Li *et al.,* 2018).

II.III.IV : Dynamique et fonction des petits ARN non codants dans les cellules germinales et la fertilité mâles

Le contenu en scnRNA a longtemps été considéré comme un reliquat de la spermatogénèse inutile pour les fonctions exercées par le spermatozoïde mature. Cependant, contrairement à la méthylation de l'ADN, le contenu en sncRNA du spermatozoïde évolue au cours de son trajet dans la sphère génitale mâle (Chu *et al.*, 2019; Sellem *et al.*, 2021). Cette évolution est caractérisée par une diminution de la

représentativité de certaines familles de scnRNA et par le gain d'autres familles, en particulier au moment du transit épidydimaire, suggérant un rôle des scnRNA ainsi acquis dans des étapes post-spermatogenèses décisives pour la fertilité mâle (maturation épidydimaire, fécondation ou développement précoce).

Si l'on s'intéresse au contenu des spermatozoïdes en scnRNA au niveau du parenchyme testiculaire, c'est-à-dire juste après la fin de la spermatogénèse et avant leur passage dans l'épididyme, on peut remarquer que la famille majoritaire des scnRNA représentés sont les piRNA (Sellem *et al.*, 2021). Cette observation est cohérente avec la fonction des piRNA dans la différenciation des cellules germinales mâles et dans la spermatogénèse. En effet, au cours de la différenciation des cellules germinales mâles, lors de la méthylation *de novo* du génome, les piRNA permettent la méthylation de certaines séquences transposables (Kuramochi-Miyagawa *et al.*, 2008). L'inactivation de MIWI2 provoque ainsi un blocage méiotique, suggérant également un rôle des piRNA dans la régulation de la méiose (Carmell *et al.*, 2007).

Cependant, au moment de la maturation des spermatozoïdes dans l'épididyme, une diminution de la proportion de piRNA et une augmentation de la proportion de miRNA, rsRNA et de tsRNA sont observées, provenant d'échanges avec les épididymosomes qui sont des microvésicules sécrétées par l'épithélium de l'épididyme (Peng *et al.*, 2012; Nixon *et al.*, 2015; Sharma *et al.*, 2016, 2018). Au moment de la fécondation les sncRNA spermatiques rencontrent avec les ARNm stockés dans l'ovocyte (Ostermeier *et al.*, 2004). Les sncRNA spermatiques pourraient donc avoir un rôle au cours de la reprogrammation épigénétique ainsi qu'au moment de la mise en route du génome embryonnaire avec la transcription de nouveaux ARNm, en régulant par exemple la stabilité des ARNm maternels ou la traduction des ARNm embryonnaires, mais également en contrôlant la réactivation des transposons et la mise en place de l'hétérochromatine embryonnaire dans un contexte de déméthylation du génome (Chen *et al.,* 2016). Cela suggérerait un rôle de ces sncRNA nouvellement acquis dans le développement. En accord avec cette hypothèse, il a été démontré qu'en traitant des spermatozoïdes

matures avec de la RNase, 90% de leur contenu en ARN était détruit. Cela a pour conséquences de diminuer les taux de développement des embryons fécondés, comparé à des conditions contrôles (Guo *et al.*, 2017). Une étude a également montré qu'en utilisant des spermatozoïdes avant leur maturation épididymaire pour réaliser des fécondations par « intra cytoplasmic sperm injection » (ICSI), afin de pallier l'absence de maturation des spermatozoïdes, il n'y avait pas naissance de souriceaux comparé à des conditions contrôles (Conine *et al.*, 2018). Pour aller plus loin cette même équipe a isolé les sncRNA des épididymosomes pour complémenter les spermatozoïdes non matures, ce qui a permis de restaurer le taux de développement, montrant l'importance des sncRNA pour le développement (Conine *et al.*, 2018). Néanmoins il existe une controverse à ce sujet. En effet dans une autre étude, des spermatozoïdes issus de la queue ou de la tête de l'épididyme ont dans les deux cas la capacité de générer une portée après une ICSI (Wang *et al.,* 2020). Cette étude ne remet pas en cause l'importance des scnRNA spermatiques pour l'embryon, mais montre néanmoins que dan certaines conditions, les miRNA acquis au cours du transit epididymaire ne sont pas indispensables pour le développement embryonnaire et fœtal. Enfin, certains tsRNA sont exprimés différentiellement dans les spermatozoïdes de patients, en fonction des résultats obtenus en FIV (Hua *et al.*, 2019).

L'ensemble de ces résultats démontre l'importance des sncRNA dans la fertilité mâle, que ce soit au niveau des étapes de spermatogénèse et de développement embryonnaire.

Conclusion

L'intégralité de ces éléments démontre l'importance de la contribution de l'épigénétique dans la fertilité mâle, ce qui fait des marques et mécanismes épigénétiques de potentielles sources de biomarqueurs. Cependant comme mentionné dans la première partie de cette introduction, ce ne sont pas les seules sources pertinentes d'information pour caractériser la fertilité mâle. Bien qu'il n'existe pas d'évaluation génomique de la fertilité mâle, la contribution génétique à la fertilité n'est en effet pas nulle. Certes les paramètres spermatiques ne sont pas des prédicteurs très précis, mais ils permettent néanmoins d'expliquer une partie de la variation phénotypique de la fertilité mâle. Ainsi

ne s’intéresser qu’à une seule catégorie de données biologiques ne fournira qu’un éclairage partiel du phénotype multi-factoriel que représente la fertilité. Dans ce contexte, l’intégration de différents types de données biologiques hétérogènes pourrait permettre d’améliorer la qualité de prédiction des modèles ainsi que de mieux comprendre comment s’élabore un phénotype tel que la fertilité mâle.

III Les méthodes d'intégration de données

L'intégration de données est un type de méthodologie ayant pour objectif d'utiliser plusieurs sources d'informations, souvent de natures hétérogènes, dans la compréhension d'un phénomène. En effet, n'utiliser qu'une seule source d'information peut être sous optimal pour l'étude de phénomènes à causes multiples. Pour illustrer ces propos, on peut prendre en exemple le lien qui existe entre la concentration en ARNm et la concentration en protéines. Bien qu'il existe un lien logique entre la présence d'un ARNm et la traduction en protéines, la corrélation entre ces deux entités biologiques est moyenne, par exemple chez la levure elle est de 0,39 (Gygi *et al.*, 1999). Ainsi n'analyser que la concentration en ARNm afin d'en déduire la concentration en protéines n'est pas inintéressant (il y a quand-même une corrélation positive) mais reste limité. Biologiquement, cela est dû au fait qu'il y a d'autres facteurs biologiques impactant cette relation comme, par exemple, la présence de miRNA, la saturation des ribosomes, la vitesse de dégradation des protéines, etc. (Maier *et al.,* 2009). Ainsi, afin de mieux comprendre la relation pouvant exister entre ARNm et protéines il est nécessaire de prendre en compte les informations ayant pour conséquence la régulation post-transcriptionnelle des ARNm, la régulation de la traduction ainsi que la dynamique des protéines traduites, autrement dit, il faut faire une intégration de ces différentes données. Comme expliqué précédemment, la fertilité mâle est un phénotype multifactoriel. Ainsi ne prendre en compte qu'un seul type de données, à l'image de l'exemple ci-dessus, n'est pas inintéressant mais reste cependant limité pour expliquer une grande partie de ce phénotype. Au cours de mon travail de thèse, j'ai donc utilisé certaines approches intégratives afin de décrire la fertilité mâle.

L'objectif de cette partie est d'exposer les différents types d'analyses intégratives et de présenter en particulier les approches utilisées dans ce manuscrit.

Les analyses intégratives ont commencé à devenir centrales en biologie depuis l'apparition des techniques de séquençage haut débit. Ces techniques permettent d'acquérir un grand nombre de données sur un même échantillon biologique. Ces données étant très souvent de grande dimension,

elles nécessitent également des outils permettant de les analyser. En s'intéressant à la littérature sur les analyses intégratives de données, on peut discerner deux grands types de méthodologies (Tini *et al.*, 2019; Eicher *et al.*, 2020; Subramanian *et al.*, 2020; Picard *et al.*, 2021)

(i) Approches exploratoires, analyses non supervisées destinées à l'analyse globale des tables étudiées, ainsi qu'à l'étude des interactions ou corrélations qui peuvent exister entre les différentes tables de données (partie III.I).

(ii) Approches prédictives, analyses supervisées destinées à prédire un phénotype et identifier des biomarqueurs prédictifs (partie III.II).

III.I : Approches exploratoires

Cette première famille de méthodes intégratives permet d'explorer un jeu de données pour en tirer des propriétés très générales, par exemple la ressemblance entre les individus ou l'identification des facteurs de variation. En général, ces méthodes sont non supervisées, c'est-à-dire que l'analyse ne prend pas en considération que les échantillons appartiennent à des catégories (exemple : échantillons contrôles contre échantillons pathologiques). Une des méthodes d'intégration permettant de réaliser ces analyses est l'Analyse Factorielle Multiple (AFM) (Escofier and Pages 1994; Abdi *et al.;* 2013). Cette méthodologie est une extension de l'Analyse en Composantes Principales (ACP), mais à la différence de cette dernière, elle est adaptée à des contextes multi-omiques.

III.I.I : L'Analyse Factorielle Multiple

L'AFM est utilisée dans le cas où les mêmes individus sont caractérisés par plusieurs variables appartenant à des groupes différents ; dans notre cas précis, il s'agit des différentes entités biologiques mesurées. Pour la suite des explications, on notera $\boldsymbol{X_1, X_2, ..., X_K}$ les matrices de données collectées pour chaque table. Chaque matrice est décrite par I individus (identiques pour chaque matrice) et J variables (différentes entre chaque table de données).

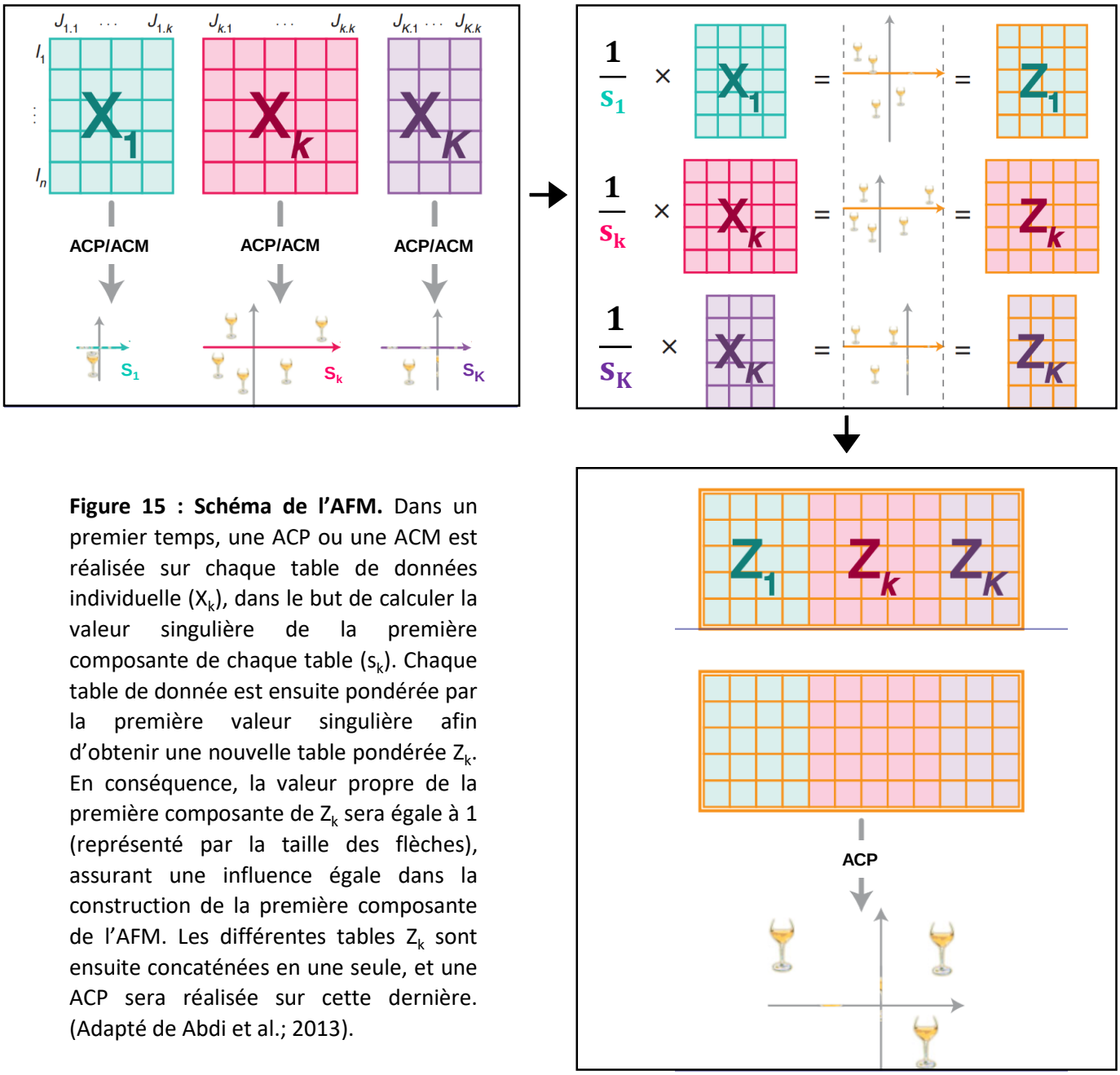


Figure 15 : Schéma de l'AFM. Dans un premier temps, une ACP ou une ACM est réalisée sur chaque table de données individuelle (X_k), dans le but de calculer la valeur singulière de la première composante de chaque table (s_k). Chaque table de donnée est ensuite pondérée par la première valeur singulière afin d'obtenir une nouvelle table pondérée Z_k. En conséquence, la valeur propre de la première composante de Z_k sera égale à 1 (représenté par la taille des flèches), assurant une influence égale dans la construction de la première composante de l'AFM. Les différentes tables Z_k sont ensuite concaténées en une seule, et une ACP sera réalisée sur cette dernière. (Adapté de Abdi et al.; 2013).

L'idée générale de l'AFM est de pondérer les différentes tables de données, de sorte à ce qu'elles contribuent de manière équivalente dans une analyse factorielle (Figure 15). L'AFM se décompose en trois étapes différentes :

La première étape consiste à réaliser une ACP, dans le cadre de données quantitatives, ou une Analyse des Correspondances Multiples (ACM) dans le cadre de données qualitatives, sur chaque table de données séparément. Elle a pour objectif de calculer la valeur propre de la première dimension de l'ACP (λ_k^1, avec λ^1 la valeur propre de la première dimension de l'ACP et k la table de données analysée) associée aux $\boldsymbol{X}_k$. Les valeurs propres représentent la variance de $\boldsymbol{X}_k$ expliquée par chaque dimension de l'ACP, ainsi la valeur propre de la première dimension de l'ACP décrit la source principale de variation. De cette valeur propre est extraite une valeur singulière calculée par : $s_k^1 = \sqrt{\lambda_k^1}$ avec s_k^1 la valeur singulière de la première dimension de $\boldsymbol{X}_k$.

La seconde étape a pour objectif de pondérer chaque table de données par leur première valeur singulière calculée à l'étape précédente. Cette étape est nécessaire pour que chaque table contribue de manière égale dans le calcul de la première dimension de l'analyse factorielle. Ainsi, on calcule :

$$Z_k = \frac{X_k}{s_k^1}$$

Avec $\boldsymbol{Z}_k$ la matrice pondérée de $\boldsymbol{X}_k$. A l'issue de cette pondération, les valeurs propres de la première dimension des différentes matrices pondérées seront toutes égales à 1, et assureront une contribution égale de chaque tableau dans l'analyse factorielle.

Enfin, la dernière étape consiste à concaténer les différentes matrices $\boldsymbol{Z}_k$ en une seule matrice $\boldsymbol{Z}$, sur laquelle sera réalisé une ACP.

Les sorties graphiques de l'AFM sont très semblables à celles obtenues dans le cadre de l'ACP. Il sera ainsi possible de projeter les individus dans les différentes dimensions, d'identifier les variables corrélées aux différentes dimensions, etc.

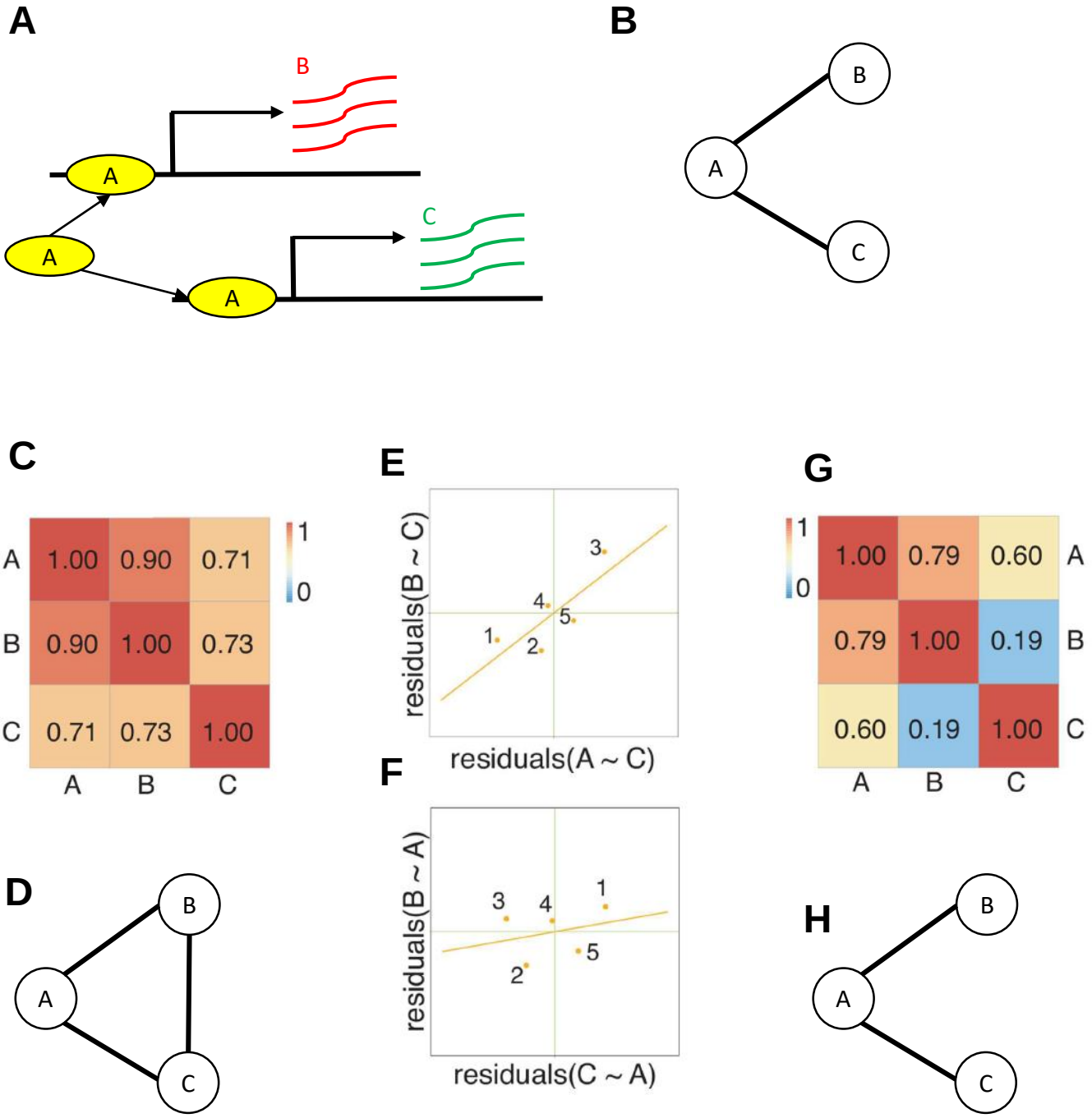


Figure 16 : Illustration de réseaux, et concept de corrélation partielle. A : Exemple biologique où la fixation du facteur de transcription A, entraîne la transcription des gènes B et C. B : Le réseau construit aura donc 3 nœuds et 2 arcs reliant A à B et B à C. C : Si l'on construit une matrice de corrélation des données de A, B et C on s'apercevra que ces trois entités sont fortement corrélées. D : Ainsi, si l'on construisait le réseau associé à partir de la matrice de corrélation, on observerait que chaque nœud est connecté à tous les autres, ce qui n'est pas le réseau théorique que nous devrions obtenir. E et F : Il faut donc travailler sur les corrélations partielles, c'est à dire la corrélation entre deux variables après avoir exclu l'influence des autres. E : Dans le cas où l'on veut analyser la relation entre A et B, il faut donc extraire l'influence C (par les résiduelles d'A et B régressées par rapport à C), puis calculer la corrélation entre ces résiduelles ; puis appliquer la même démarche à A et C puis à B et C. G : En faisant cela, on observe que la corrélation partielle de B et C est faible. H : Ainsi, lorsque l'on construit de réseaux, il n'y a plus d'arc entre B et C. (Adapté de Hawe et al. ; 2019).

Il existe bien évidemment d'autres méthodes permettant de répondre à ce genre de problématiques ou de questions. Il y a par exemple la méthode Multi-Omics Factor Analysis (MOFA) se basant sur la même stratégie que l'AFM : explorer les sources de variations importantes existant entre les jeux de données (Argelaguet *et al.*, 2018). On peut également noter des méthodes de classification ayant pour objectifs d'identifier les ressemblances entre individus comme par exemple iCluster (Shen *et al.,* 2009).

III.I.II : Etude des relations entre variables

Un autre aspect intéressant dans les analyses intégratives est de pouvoir appréhender les relations de corrélation pouvant exister entre les variables composant un jeu de données. Cela peut se faire au sein d'une table de données, ou également entre tables de données. Encore une fois, il existe différentes méthodes permettant de réaliser ce travail. Il y a par exemple les méthodes de l'analyse des corrélations canoniques (CCA) et la régression des moindres carrés partielles (PLS) (Hotelling, 1992; Wold *et al.,* 2001). Ces deux méthodes sont proches des ACP et sont utilisées quand deux jeux de données différents ont été acquis sur les mêmes individus. L'objectif ici n'est pas de construire des composantes maximisant la variance, mais des composantes maximisant la corrélation dans le cas des CCA ou la covariance dans le cas de la PLS entre les deux jeux de données analysés. Les premières composantes étant construites à partir des variables ayant la plus grande corrélation ou covariance entre jeux de données, cela permet de faire apparaître des relations entre variables.

III.I.II.I : L'inférence de réseaux

L'inférence de réseaux a pour objectif de mettre en évidence les relations entre les variables, ainsi que de représenter graphiquement les résultats (Figure 16). Dans ces représentations un nœud symbolise une variable, et un arc symbolise une liaison entre deux nœuds, autrement dit entre deux variables. Ces réseaux peuvent être très utiles en biologie pour représenter par exemple la corrélation de facteurs de transcription ou de marques épigénétiques à la transcription en ARN ou la traduction en protéines.

Soit $\boldsymbol{X}$ un tableau de données décrit par I individus et P variables. Pour représenter le réseau d'interactions des P variables de $\boldsymbol{X}$, on estime une matrice $\boldsymbol{M}$ de dimension P x P, permettant de synthétiser l'information d'interaction pour toutes les variables de notre jeu de données. Chaque cellule de cette matrice reflète la relation existant entre une variable i et une autre variable j. Si $\boldsymbol{M}_{ij} = 0$, cela veut dire que les deux variables n'interagissent pas ; si l'on reprend l'exemple précédent, cela se traduit par le fait que la présence du facteur de transcription i n'est pas corrélée à la présence de l'ARN j. A l'inverse si $\boldsymbol{M}_{ij} \neq 0$, cela veut dire qu'il existe un lien entre i et j.

Les méthodes d'inférence de réseaux ont donc pour objectif de calculer la matrice $\boldsymbol{M}$ mettant en évidence les variables liées. Il existe différentes méthodes permettant de travailler dans le cadre de données omiques, et deux d'entre elles sont présentées ci-dessous. La liste n'est cependant pas exhaustive, et les autres types de méthodes sont décrites dans la revue suivante (Hawe *et al.,*2019).

Une des méthodes de référence est le Graphical Lasso, elle appartient au domaine des modèles graphiques gaussiens (Friedman *et al.,* 2008). Les modèles graphiques gaussiens sont utilisés lorsque les *P* variables d'un jeu de données $\boldsymbol{X}$ suivent une loi normale. Dans le cas des modèles graphiques gaussien, la matrice $\boldsymbol{M}$ que l'on a présenté précédemment est calculée en inversant la matrice de covariance de $\boldsymbol{X}$; elle est appelée matrice de précision et est souvent représentée de cette façon : Σ^{-1}. L'avantage de cette matrice, comparé à la matrice de covariance, est qu'elle met en évidence les variables en corrélation partielle (Figure 16) Ainsi, au sein de cette matrice, si $\Sigma^{-1}_{ij} = 0$ cela veut dire que les variables sont conditionnellement indépendantes, et à l'inverse si $\Sigma^{-1}_{ij} \neq 0$ alors i et j sont conditionnellement dépendantes. En calculant la matrice de covariance de X il est donc théoriquement possible de calculer Σ^{-1} et de représenter le réseau associé. Cependant, dans le cadre des données omiques, c'est-à-dire quand I << P, il n'est pas possible d'inverser la matrice de covariance associée. Il faut donc procéder à une estimation de Σ^{-1}. Pour simplifier la lecture dans la littérature, la matrice Σ^{-1} est noté Ω et nous ferons de même pour la suite des explications. Dans le Graphical Lasso les paramètres de la matrice Ω sont estimés de cette façon :

$$\Omega = argmax(L(\Omega) - \lambda\|\Omega\|_1) \; avec \; \|\Omega\|_1 = \sum |\Omega_{ij}|$$

Avec L le log de la vraisemblance des modèles graphiques gaussiens, $\lambda\|\Omega\|_1$la pénalité de la norme L1 et λ le paramètre de régularisation. L'introduction de la pénalité de type L1 permet aux interactions faibles entre i et j d'être égales à 0. La matrice Ω est par conséquent qualifiée de « sparse » (parcimonieuse), et plus le paramètre λ est élevé plus le nombre de paramètres égaux à 0 sera grand, et plus la matrice sera « sparse ». Cette méthodologie permet donc de travailler dans le cadre de données -omiques. Cependant, les modèles graphiques gaussiens supposent que les données analysées suivent une loi normale, ce qui n'est pas toujours le cas dans le cadre d'analyses intégratives de données hétérogènes. Des méthodes comme les Modèles Graphiques Mixtes ont permis d'intégrer des données qualitatives, mais suppose néanmoins que les données quantitatives soient gaussiennes (Lee and Hastie, 2015).

Afin de prendre en considération des données avec des distributions hétérogènes, des méthodes comme celle présentée dans le package R GENIE3 ont vu le jour, et ne supposent pas de distribution particulière des variables de $\boldsymbol{X}$(Huynh-Thu *et al.,* 2010). Cette méthode consiste à utiliser un modèle dans le but régresser la variable p_i contre toutes les autres ; et de procéder ainsi pour toutes les variables de $\boldsymbol{X}$. Le modèle de régression utilisé sont les forêts aléatoires (détaillées dans le chapitre suivant). Elles permettent de calculer l'importance relative des variables dans la construction d'un modèle. Plus une variable p_j a une valeur d'importance relative élevée dans la régression de p_i, alors plus p_j est liée à p_i. En appliquant cette méthode, on obtient une matrice $\boldsymbol{M}$ de dimension $P \; x \; P$, ou chaque cellule p_{ij} contient l'importance relative des variables obtenue lors de la régression de p_j par les autres variables. Après définition d'un seuil, on peut s'appuyer sur cette matrice dans le but de construire un réseau.

III.II : Prédiction des phénotypes

Au cours de ce dernier chapitre sur les méthodes intégratives, nous allons aborder les modèles de prédiction. Ces méthodes ont pour objectif d'établir une relation entre une variable à expliquer y et des variables explicatives x. La nature de la relation dépend du type de méthodologie utilisé. Dans le cadre de ce travail de thèse, la variable y est la fertilité des animaux (qualitative à deux modalités : fertile ou subfertile) et les prédicteurs sont les différentes variables mesurées pouvant appartenir aux données de méthylation de l'ADN, des sncRNA, etc. Cette partie du manuscrit sera donc consacrée à exposer les méthodes utilisées au cours de ce travail.

III.II.I : Régression logistique avec pénalité Lasso

Les régressions logistiques sont des méthodes utilisées lorsque la variable à expliquer y est de nature qualitative à deux modalités, les variables explicatives (x) pouvant être quantitatives ou qualitatives. La relation décrite ci-dessous permet de connaître la probabilité qu'un individu i appartiennent à la classe 1 en fonction des différentes valeurs de x.

$$P(1|X) = \frac{e^{b_0+b_1x_1+\cdots+b_jx_j}}{1+e^{b_0+b_1x_1+\cdots+b_jx_j}}$$

Chaque valeur de x est associée à un coefficient b, ce qui permet d'établir la relation pouvant exister entre les prédicteurs et la variable à prédire. Les coefficients sont calculés de façon à maximiser la vraisemblance. De cette façon, les coefficients calculés sont ceux décrivant de manière optimale la relation entre x et y. Dans ce travail de thèse, nous sommes cependant dans un cadre où des données -omiques sont analysées. Le nombre de variables étant supérieur au nombre d'individus, nous avons utilisé un critère de pénalité de type Lasso pour le calcul des coefficients b_j, en permettant aux coefficients associés à des variables peu explicatives de la variable y d'être fixés à 0.

III.II.II : Forêts aléatoires et Gradient Boosting

Les forêts aléatoires, ainsi que le gradient boosting, sont des méthodes qualifiées « d'ensemble », c'est-à-dire qu'elles n'utilisent pas un modèle de prédiction unique mais plusieurs modèles pour bâtir la prédiction (Breiman, 2001; Friedman, 2001). Dans ces deux cas, le modèle utilisé est le même et s'agit d'un arbre de décision (Breiman *et al.*, 2017).

III.II.II.I : Arbres CART

Les arbres de décision sont des outils mathématiques permettant de faire de la régression (prédiction de variables quantitatives) ou de la classification (prédiction de variables qualitatives). Ces arbres sont constitués de nœuds (test sur une variable) et de feuilles (il s'agit des nœuds terminaux, n'effectuant pas de test et reflétant la prédiction). Dans le cadre de problèmes de régression les feuilles sont constituées par un nombre, alors que dans les problèmes de classification il s'agit d'une classe.

Soit $\boldsymbol{X}$ un jeu de données décrit par I individus et P variables. Les arbres de décisions sont construits de sorte à ce qu'ils réduisent au plus l'erreur de classification. Pour cela, deux éléments doivent être identifiés : la variable sur laquelle sera effectuée le test ainsi que son point de coupure. Pour cela l'intégralité des *P* variables ainsi que des points de coupure seront testés, les deux paramètres choisis seront ceux minimisant les erreurs, et la métrique pour qualifier l'erreur dépend du problème considéré. Dans le cas de problèmes de régression ce sera la minimisation de l'erreur quadratique. Dans le cas de problèmes de classification il s'agira de minimiser le critère de Gini, un indice d'impureté : plus ce critère est faible plus les feuilles sont homogènes en terme de classe. Ces étapes sont répétées jusqu'à ce que l'arbre final soit construit.

Cependant, les arbres décisionnels sont de mauvais prédicteurs car ils ont une grande variance. C'est-à-dire que les performances d'un arbre de classification sont peu reproductibles, la robustesse de ces méthodes est donc assez faible. Une façon d'augmenter la robustesse de ces méthodes est d'accroître

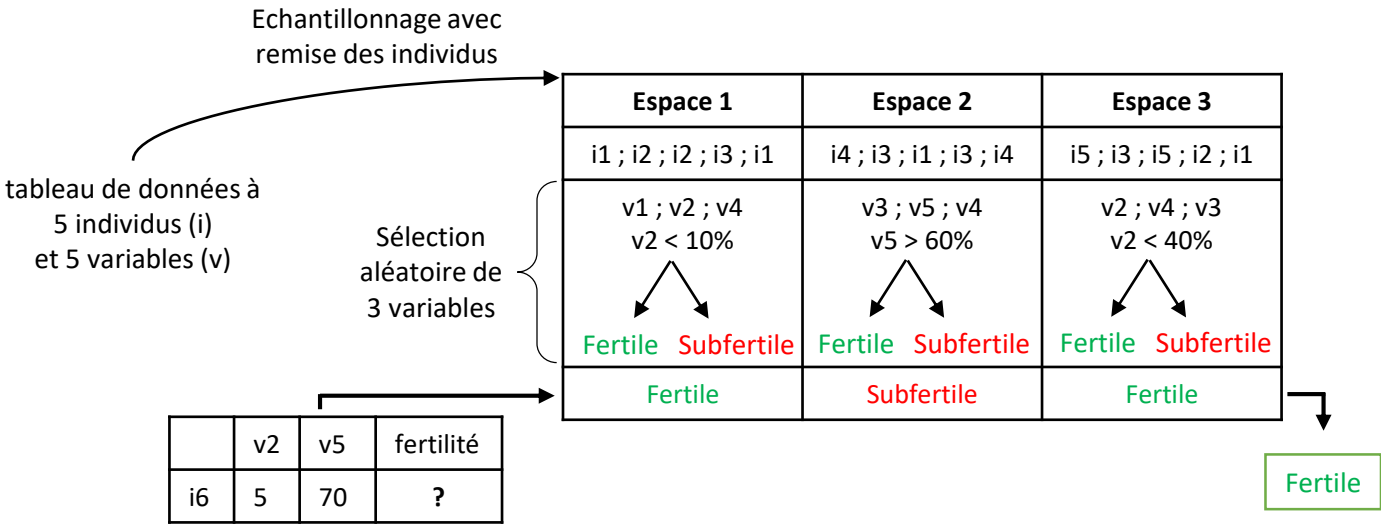


Figure 17 : Exemple de construction d'une forêt aléatoire avec un jeu de données constitué de 5 individus et de 5 variables. Dans un premier temps, chaque espace (3 dans ce cas) est construit en réalisant un tirage aléatoire avec remise des individus (« bootstraping »). Nous construisons ensuite un arbre par espace ; pour l'exemple les arbres sont petits avec un seul nœud par arbre. Avant de définir la variable sur laquelle sera fait le test, nous faisons un tirage aléatoire des variables. Par exemple, dans l'espace 1, v1 v2 et v3 sont tirées, cela veut dire que l'arbre ne pourra se construire qu'en prenant ces variables en considération. Dans l'espace 1 par exemple, l'arbre prend appui sur la variable v2 avec le critère « <10% » pour classifier les animaux. Lors de la prédiction d'un nouvel individu dont on ne connaît pas la prédiction (dans l'exemple i6), on le classifiera par l'ensemble des arbres construits. Ici deux arbres sur trois l'on classifié comme « Fertile », ce sera donc la classe prédite pour cet individu par les forêts aléatoires.

le nombre d'arbres décisionnels générés. Les méthodes des forêts aléatoires et du gradient boosting implémentent cette stratégie et sont décrites ci-dessous.

IIII.II.II.II : Forets aléatoires

Afin de générer une grande quantité d'arbres de décision à partir d'un seul jeu de données, l'idée générale des forêts aléatoires est de construire différents espaces, représentés par différents individus et différentes variables, afin de générer un arbre par espace (Figure 17). En prenant ensuite en compte l'information apportée par tous les arbres, une information de prédiction unique sera obtenue pour chaque individu. Comme les arbres CART peuvent travailler sur des problèmes de régression et de classification, les forêts aléatoires partagent également cette propriété.

L'algorithme commence par initialiser K espaces différents, où un arbre sera généré par espace. Ces espaces sont générés en réalisant un « bootstrap » des individus, ce qui signifie un tirage aléatoire avec remise. Chaque espace sera donc constitué d'individus différents, et chaque individu pourra être représenté plusieurs fois au sein d'un espace. Au sein de chaque espace sera généré un arbre de décision. Dans la construction des arbres CART, nous avons vu que l'algorithme utilisait toutes les variables explicatives disponibles dans le but d'identifier la variable et le point de coupure utilisés pour les différents nœuds de l'arbre. Cette méthodologie est modifiée dans la construction des forêts aléatoires car avant la construction d'un nœud, un sous-échantillon réduit des variables explicatives est extrait, contraignant ainsi la construction du nœud parmi les variables tirées dans ce sous-échantillon. Cette procédure sera réalisée pour la construction de chaque nœud de chaque arbre des différents espaces. De par ces deux propriétés les arbres générés seront hétérogènes entre chaque espace.

Lors de la prédiction d'un nouvel individu, la moyenne obtenue par les K arbres sera calculée dans le cadre d'un problème de régression. Dans le cadre d'un problème de classification, il s'agira de la classe majoritaire présentée par les K arbres qui permettront d'établir la prédiction.

Jeu de données

	X1	X2	y
i1	25	13	93
i2	27	12	85
i3	89	14	50
i4	89	12	49

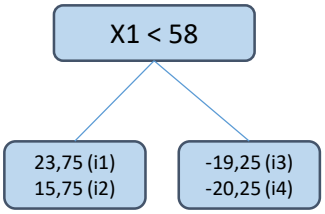
Etape 1

	y	Fo(x)
i1	93	69,25
i2	85	69,25
i3	50	69,25
i4	49	69,25

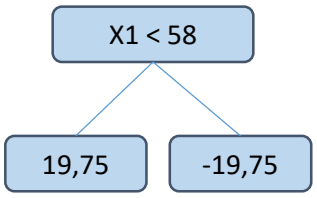
Etape 2.1

	y	Fo(x)	R1
i1	93	69,25	23,75
i2	85	69,25	15,75
i3	50	69,25	-19,25
i4	49	69,25	-20,25

Etape 2.2



Etape 2.3



Etape 2.4

F1(x) = Fo(x) + 0,1(Rjm)

	y	Fo(x)	F1(x)
i1	93	69,25	71,225
i2	85	69,25	71,225
i3	50	69,25	67,325
i4	49	69,25	67,325

Figure 18 : Exemple de construction d'un Gradient Boosting, dans un problème de régression, avec un jeu de données composé de 4 individus de deux variables explicatives (x1 et x2) et d'une variable à prédire (y). Etape 1 : la première étape de l'algorithme consiste à réaliser une prédiction unique pour tous les individus. Dans les problèmes de régression, il s'agit de la moyenne de (y) pour l'ensemble des individus. L'étape 2.1 consiste à calculer la pseudo-résiduelle (R1 dans l'exemple) pour chaque individu. Il s'agit de la différence entre la valeur à prédire (y) et la valeur de prédiction la plus récente (dans l'exemple Fo(x)). La pseudo-résiduelle peut être vue comme ce qu'il reste à prédire avant d'obtenir le bon résultat. L'étape 2.2 consiste à construire un arbre permettant de minimiser l'erreur de prédiction de la pseudo-résiduelle. On ne cherche pas à prédire (y) mais (R1). La variable X1 est utilisée pour la construction de l'arbre car elle permet de minimiser l'erreur de prédiction de R1. Dans l'étape 2.3, on calcule une valeur de feuilles unique étant donné qu'il y a plusieurs valeurs dans chacune d'entre elles, qui correspond à la moyenne des valeurs au sein de chaque feuille. L'étape 2.4 permet de calculer la nouvelle prédiction pour chaque individu, en additionnant la dernière prédiction avec la nouvelle (valeur de la feuille dans l'arbre en fonction du résultat du test sur le nœud). Par exemple i1 a une valeur Fo(x) de 69,25 ; sa valeur x1 est de 25 l'arbre prédit donc la pseudo-résiduelle à 19.75 ; F1(x1) = 69,25 + 0.1(19,75) = 71,225. La valeur « 0.1 » est le taux d'apprentissage qui permet de contrôler la vitesse à laquelle converge l'algorithme. Pour tous les individus on observe que la valeur de F1(x) est plus proche que F0(x) de (y), ce qui montre que cette étape a permis de se rapprocher de la vraie valeur. L'étape 2 est contenue dans une boucle qui est répétée un nombre *M* de fois, valeur fixée par l'utilisateur. Rjm : valeur de la pseudo-résiduelle calculée pour la feuille j à l'étape m.

III.II.II.III : Gradient boosting

La deuxième méthode d'ensemble basée sur les arbres décisionnels est le gradient boosting, qui diffère totalement dans la philosophie de construction par rapport aux forêts aléatoires. Dans les forêts aléatoires les arbres sont construits de manière indépendante les uns des autres, ce qui n'est pas le cas dans le gradient boosting. En effet, l'idée générale du gradient boosting est de générer des arbres qui vont successivement expliquer l'erreur non expliquée des arbres précédents. Ainsi, chaque arbre n (à l'exception du premier) est construit par rapport aux résultats de l'arbre (n-1). L'algorithme du gradient boosting est détaillé dans le livre de Hastie *et al.*, 2001. Il se décompose en deux étapes principales (Figure 18).

La première étape de l'algorithme a pour but de l'initialiser en réalisant une prédiction unique pour tout les individus. Dans le cas d'une variable quantitative, par exemple, il s'agit de construire un arbre avec une seule feuille, autrement dit, tous les individus auront la même prédiction, qui correspond à la moyenne de la variable y (variable à prédire).

La deuxième étape va permettre de construire les arbres, et est insérée dans une boucle de 1 à M, autrement dit, elle se terminera quand M arbres (valeur fixée par l'utilisateur) seront construits.

On note y la variable à prédire et x les variables explicatives. On choisit une fonction de coût L, en fonction du problème considéré. On note $F_{m-1}(x)$ la prédiction obtenue à l'étape (m-1). On calcule tout d'abord une « pseudo-résiduelle », qui dans le cas d'un problème de régression correspond, pour chaque individu, à l'écart entre la valeur prédite par l'arbre précédent et la valeur vraie. La pseudo-résiduelle r_{im}, pour l'individu i à l'étape m, est obtenue par la formule suivante :

$$r_{im} = -\left[\frac{\partial L\big(y_i, F_{m-1}(x_i)\big)}{\partial F_{m-1}(x_i)}\right]$$

On va ensuite construire un arbre de prédiction avec pour objectif non pas de prédire la valeur de y, mais la pseudo-résiduelle calculée précédemment. Cette étape est une différence fondamentale entre le gradient boosting et les forêts aléatoires puisque l'on cherche ici à expliquer les écarts de prédiction

obtenus par les arbres précédents. On actualise ensuite la prédiction en additionnant les résultats obtenus par le nouvel arbre créé aux prédictions précédentes. En ajoutant la nouvelle valeur, on s'approche de la vraie valeur des individus. On réitère ensuite ce processus *M* fois. Ainsi, à chaque itération, la valeur de pseudo-résiduelle va diminuer et on se rapprochera de la vraie valeur recherchée.

Lors de la prédiction d'un nouvel individu, la somme de la valeur obtenue par sa classification au sein de chaque arbre déterminera sa prédiction.

III.II.III Réseaux de neurones

Pour conclure cette partie sur les méthodes permettant de réaliser de la prédiction, voici une brève introduction aux réseaux de neurones qui sont un champ thématique très complet et complexe. Comme les deux méthodes précédentes, les réseaux de neurones sont également des méthodes d'ensemble. Toutefois, ils ne se basent pas sur des arbres décisionnels, mais sur un neurone formel décrit par la formule ci-dessous.

$$\varphi(\sum_{i=0}^{m} w_i x_i)$$

Avec φ la fonction d'activation, w_o une valeur de biais associée à une variable fictive de valeur 1, w_i les coefficients du modèle associés aux variables x_i. Dans la parenthèse, on retrouve une expression similaire à ce qui est trouvé dans les régressions linéaires multiples, chaque variable x_i est associée à un poids w_i, avec la présence d'un intercept w_o. Ce qui différencie le neurone formel des régressions linéaires est que la relation est contenue dans une fonction d'activation. Dans le cas d'une régression linéaire multiple, la relation entre les y et les x est linéaire. Ainsi, en utilisant une fonction d'activation on peut transformer l'information pour ne pas être limité par une relation linéaire afin de modéliser d'autre types de relation (exponentielle par exemple). Il existe différents types de fonction d'activation, permettant de modéliser différents types de relations.

Comme les réseaux de neurones sont des méthodes d'ensemble, plusieurs neurones formels composent ce réseau. En général, les neurones sont répartis par couche et le réseau est construit de sorte à ce que les neurones de la couche (n-1) et les neurones de la couche (n+1) soient connectés aux neurones de la couche n. Démultiplier les neurones et les couches est utile afin de modéliser des relations toujours plus complexes. En somme plus le problème est complexe, plus l'architecture du réseau permettant de le modéliser doit l'être. Comme pour le neurone formel, c'est en ajustant les poids associés à chaque variable ou aux produits de neurones, que l'on peut établir la relation entre les x et y. L'estimation est réalisée en même temps pour tous les poids associés à une variable pour chaque neurone par un processus appelé rétropropagation de gradient, qui ne sera pas détaillé dans ce manuscrit, toutes les explications peuvent être trouvées dans cet article (Werbos, 1990).

Résultats

Objectifs de la thèse et structure de la partie Résultats

A la lumière des éléments exposés dans l'introduction, on comprend que le méthylome spermatique est une marque épigénétique intéressante dans le cadre de l'analyse de la fertilité mâle. Au moment de l'écriture de ce manuscrit seules 6 études ayant analysé le méthylome spermatique de taureaux fertiles et subfertiles de manière pan-génomique ont été publiées (Verma *et al.*, 2014; Kropp *et al.*, 2017; Fang *et al.*, 2019; Gross *et al.*, 2020; Narud *et al.*, 2021; Takeda *et al.*, 2021). De plus, ces études ont toutes été menées à partir d'effectifs réduits (n=3 à 9 taureaux par groupe) et sur différentes races, limitant la portée des conclusions établies et les applications sur des taureaux élevés dans des centres français. C'est pour cela qu'une grande partie de ce travail a porté sur l'analyse du méthylome spermatique de taureaux, avec des objectifs variés.

(i) La variabilité génétique chez les bovins, sur laquelle s'appuie la sélection génomique, peut affecter les patrons de méthylation de 2 manières : elle peut d'abord supprimer directement le site CpG cible sur un ou les deux allèles, par des polymorphismes touchant le C et/ou le G du CpG. Il est également possible qu'elle puisse moduler la méthylation de manière indirecte, par le biais de QTL de méthylation (sites génomiques qui affectent la méthylation à distance), comme décrit chez l'Homme (Do *et al.*, 2017). Cet aspect, qui ne peut être abordé qu'à partir de cohortes de centaines d'individus avec des données de méthylation et des génotypes disponibles, n'a pas été traité au cours de la thèse. Un premier objectif a donc été d'analyser l'impact direct de sites CpG polymorphes sur les données de méthylation de l'ADN, dans le but de définir une stratégie adéquate permettant d'analyser la variabilité épigénétique et non génétique.

(ii) Une fois la stratégie de gestion de la variabilité génétique définie, les données de méthylation ont été analysées dans le cadre de la fertilité mâle. Pour cela, deux cohortes de taureaux issus de deux races différentes (Montbéliarde et Holstein) et constituées d'animaux fertiles et subfertiles ont été analysées. L'objectif de cette analyse était d'une

part d'émettre des hypothèses biologiques quant à l'impact de différentiels de méthylation sur la fertilité, et d'autre part de construire un modèle de prédiction de la fertilité des taureaux (Costes *et al.*, 2022).

(iii) Enfin, la méthylation de l'ADN spermatique varie en fonction de l'âge chez plusieurs espèces mammifères dont le bovin (Jenkins *et al.,* 2018). Les variations sont très importantes dans la période péri-pubertaire, mais peuvent également être observées à un rythme plus lent au cours de la carrière de l'animal (Lambert *et al.*, 2018; Takeda *et al.*, 2019). Cet aspect est important dans une perspective d'application des résultats de la thèse sur le terrain. En effet les taureaux qui ont fait l'objet d'une sélection génomique sont généralement exploités dès la maturité sexuelle atteinte et la production de semence stabilisée (à partir de 13 mois environ), et ont une carrière courte (jusqu'à 24 mois en général) car de nouveaux taureaux porteurs d'une génétique plus intéressante arrivent sans cesse sur le marché. Il est donc impératif de déterminer la dynamique du méthylome spermatique au cours de la carrière, et en particulier dans les stades précoces, pour s'assurer de la stabilité des modèles de prédiction de la fertilité (dont l'intérêt est d'être utilisés tôt dans la carrière). De ce fait, la méthylation de l'ADN a également été analysée en fonction de l'âge des animaux. De plus, la fertilité des animaux peut varier au cours de leur carrière, avec la production ponctuelle de séries d'éjaculats avec une mauvaise fertilité. Dans le but de déterminer si le modèle pourrait être appliqué non seulement pour prédire la fertilité moyenne du taureau, mais également pour détecter ces déviations de fertilité, des échantillons de semence présentant des variations intra-individuelles de fertilité ont été analysés au cours de la thèse.

Ces analyses de la méthylation font l'objet de la partie I des Résultats. Par ailleurs, cette thèse s'insère dans un programme plus large : le programme SeQuaMol (« Qualité Moléculaire de la Semence bovine »), qui vise à identifier des marqueurs de la qualité de la semence au sein de données moléculaires de différentes origines. Au sein de ce programme, des données sur l'expression des

sncRNA, le lipidome, le proteome, le glycome, le ratio histone/protamine et certains paramètres spermatiques ont été produites strictement sur les mêmes échantillons de semence. De plus, les taureaux présents dans les cohortes exploitées ont également été génotypés comme tous les taureaux d'IA issus de la sélection génomique. Comme expliqué dans l'introduction, les démarches intégratives sont intéressantes lors de l'analyse de phénotypes multifactoriels. Ainsi, une première intégration des données de méthylation de l'ADN, d'expression des scnRNA, des paramètres spermatiques et des génotypes a été réalisée dans la thèse. Ces données ont été utilisées en première intention car elles ont été disponibles plus tôt au cours de la thèse. Les objectifs de cette intégration sont d'analyser le comportement des variables en lien avec la fertilité et d'évaluer dans quelle mesure l'apport de données supplémentaires pouvait permettre d'améliorer la précision des modèles (article 2). L'intégration de données fait l'objet de la partie 2 des Résultats.

Pour des raisons stratégiques, il a été décidé par l'ensemble des partenaires impliqués dans SeQuaMol que seules les données obtenues en race Montbéliarde (race française mettant en jeu des entreprises de sélection locales) seraient intégralement publiées et diffusées lors de communications dans des congrès. La publication des données obtenues en race Holstein (race de diffusion internationale, avec une forte compétition entre entreprises de sélection) ne pourra être envisagée qu'une fois la stratégie de valorisation des résultats définitivement arrêtée. Pour cette raison, les résultats obtenus en race Montbéliarde sont essentiellement décrits dans les articles, tandis que ceux obtenus en Holstein sont rédigés en français dans le corps de la thèse.

Enfin, pour faciliter la compréhension des résultats exposés dans les parties I et II, j'ai fait le choix de présenter d'abord les dispositifs expérimentaux et les types de données sur lesquelles j'ai travaillé. Cette partie ne constitue pas une partie « Matériels et Méthodes » à proprement parler (ceux-ci sont décrits dans les articles), c'est pourquoi elle a été intégrée dans les Résultats.

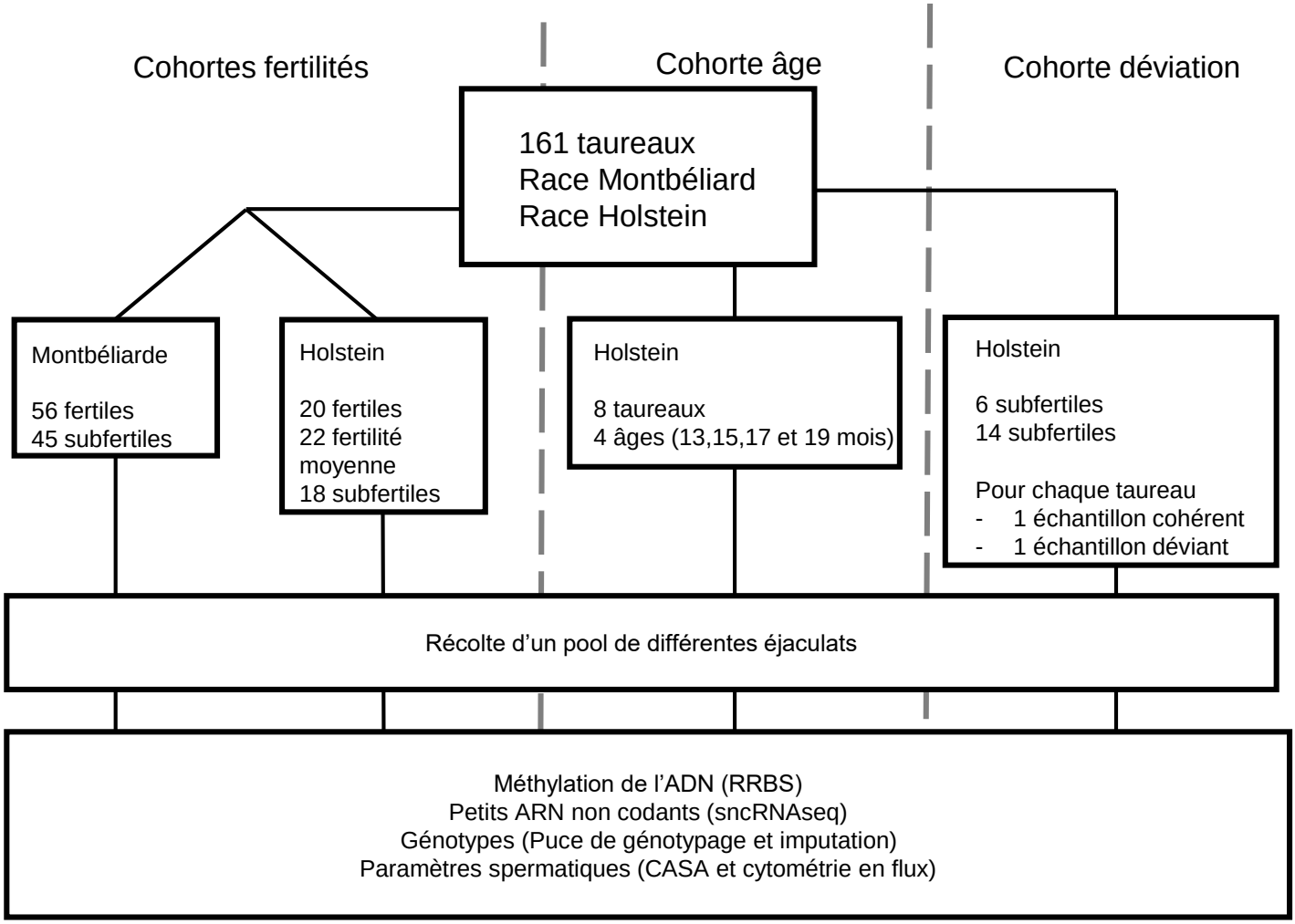


Figure 19 : Description des cohortes de taureaux. Ce travail de thèse s'appuie sur une collection de semence de 161 taureaux. Cette cohorte a été divisée en trois sous –cohortes dans le but de répondre à trois grandes thématiques. La cohorte « fertilité » est composée de deux races différentes (Holstein et Montbéliarde) et a pour objectif de comparer des taureaux fertiles et subfertiles. La cohorte « âge » ne contient que des taureaux Holstein, afin de suivre la dynamique des marques d'intérêt entre 13 et 19 mois. Enfin, la cohorte « déviation » a pour objectif d'identifier les mécanismes moléculaires associés à une déviation de la fertilité. Pour l'ensemble de ces données, la méthylation de l'ADN, les sncRNA, les génotypes et les paramètres spermatiques ont été analysés. Pour constituer chaque échantillon, plusieurs éjaculats collectés dans un intervalle de temps limité ont été réunis, afin de s'affranchir d'effets très ponctuels.

Cohorte d'animaux et données traitées au cours de la thèse

Ce travail s'est appuyée sur des cohortes et des données acquises au cours du projet SeQuaMol sur de la semence congelée et conditionnée sous forme de paillettes, commercialisée à des fins d'IA. Cette cohorte est constituée de 2 races différentes : Montbéliard et Holstein. Ces animaux ont été choisis afin de répondre à 3 problématiques et sont donc regroupés en 3 cohortes différentes : une cohorte « fertilité », une cohorte « âge » et une cohorte « déviation » (Figure 19).

Les animaux de la cohorte fertilité sont issus de deux races, 101 de la race Montbéliarde et 60 de la race Holstein. L'objectif de cette cohorte est d'explorer les différences pouvant exister entre des animaux fertiles et des animaux subfertiles mais néanmoins commercialisés et utilisés en IA. Ces animaux ont été classifiés comme étant fertiles ou subfertiles à l'aide du TNR 56 corrigé des effets environnementaux (voir introduction partie I.II.I). Avec ce phénotype, 57 animaux de race Montbéliarde ont été attribués à la classe fertile et 44 à la classe subfertile. Dans la race Holstein, 20 animaux ont été attribués à la classe fertile, 18 à la classe subfertile tandis que les 22 taureaux restants sont de fertilité moyenne. Bien que les différences de TNR 56 entre animaux fertiles et subfertiles soient hautement significatives, aucun taureau de cette cohorte n'est infertile et les écarts de TNR 56 sont faibles (Figure 20). Il a été démontré chez le bovin que le méthylome spermatique évoluait en fonction de l'âge (Lambert *et al.*, 2018; Takeda *et al.*, 2019), ainsi les éjaculats analysés ont été collectés à des âges équivalents pour tous les animaux, afin de limiter cette source de biais (entre 17 et 19 mois). Les analyses réalisées dans ce travail de thèse ont été faites dans chaque race séparément. La cohorte fertilité a été utilisée pour l'analyse de la méthylation et l'intégration de données. Dans le cadre de l'analyse de la fertilité des animaux de ces cohortes, nous nous sommes placés au niveau « taureau ». Cela signifie que les différents éjaculats caractérisant un taureau ont été sélectionnés de sorte à avoir des performances sur le terrain en adéquation avec la fertilité du taureau sur l'ensemble de sa carrière.

Les animaux de la cohorte âge sont tous de race Holstein. Cette cohorte a pour objectif d'observer la dynamique des marqueurs moléculaires analysés au cours de la vie d'un animal. Ainsi, les éjaculats de

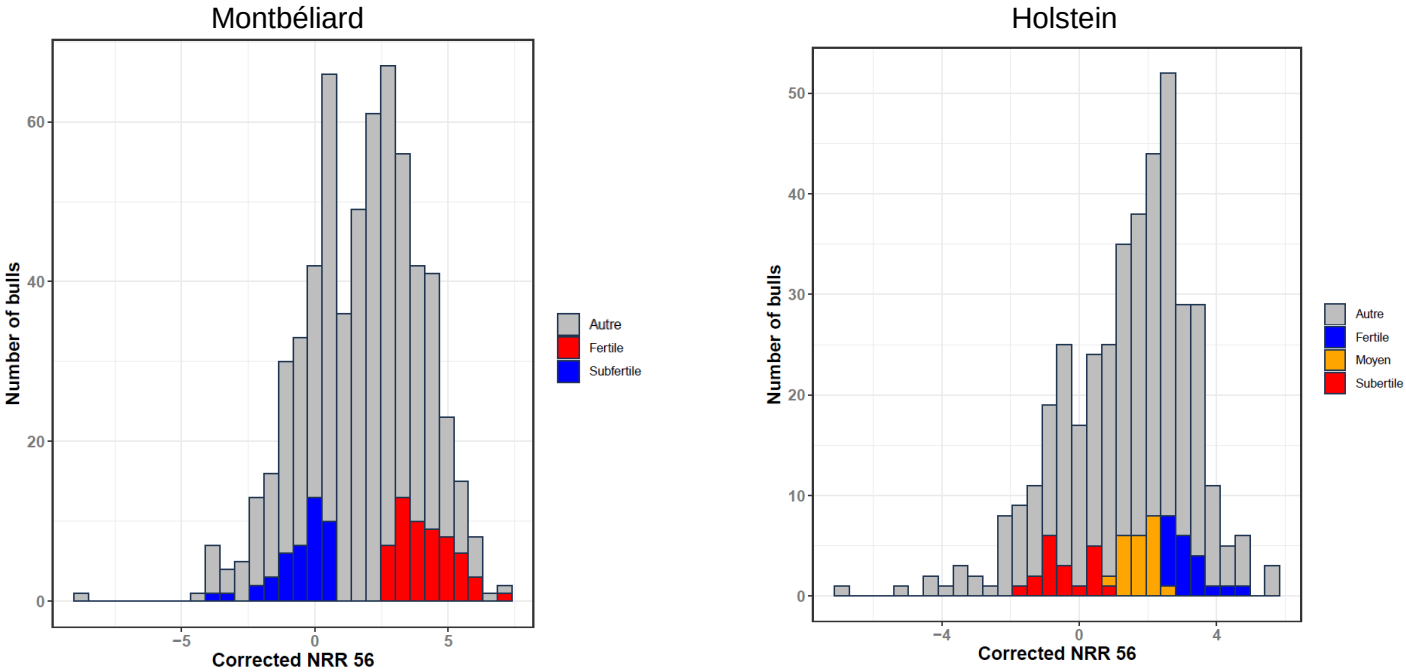


Figure 20: Performances de fertilité des animaux de la cohorte Montbélaird et Holstein. Graphique représentant sur l'axe x, les TNR 56 corrigés des taureaux, et sur l'axe y le nombre de taureaux. Les taureaux sont coloriés en fonction de leur groupe de fertilité. Les barres grises représentent les « autres » taureaux, c'est-à-dire ceux n'étant pas inclus dans la cohorte, nés dans les mêmes années que ceux présents dans la cohorte et appartenant à la même entreprise de sélection.

8 taureaux ont été suivis à 4 âges différents (13, 15, 17 et 19 mois). Cette durée couvre les âges pendant lesquels les taureaux de race Holstein sont généralement exploités afin de produire de la semence. Pour atténuer d'éventuels effets affectant des éjaculats individuels, dans la mesure du possible plusieurs éjaculats ont été réunis pour constituer chaque échantillon de cette cohorte (de 1 à 4 éjaculats, collectés dans un intervalle de moins d'un mois) (Annexe 1).

Enfin, la dernière cohorte « déviation » est également constituée d'animaux de race Holstein uniquement. Au cours de la carrière d'un taureau il n'est pas rare d'observer des déviations de fertilité. Ce serait le cas d'un taureau que l'on pourrait qualifier de fertile, et qui brutalement aurait une baisse de fertilité due à un problème sanitaire ou à un stress thermique par exemple. Cette cohorte a donc été construite en réunissant pour chaque animal, des éjaculats « cohérents » et des éjaculats « déviants » par rapport à sa fertilité. Les échantillons « cohérents » et « déviants » ont été collectés dans un laps de temps de moins d'un mois pour chaque paire d'échantillon. Ainsi, 6 animaux subfertiles et 14 animaux fertiles sont chacun caractérisés par un échantillon subfertile et un échantillon fertile (Annexe 2).

Au vu de la dimension importante des cohortes présentées ci-dessus, la décongélation des paillettes, la constitution des échantillons de semence, la mesure des paramètres spermatiques et l'extraction des acides nucléiques ont dû obligatoirement être effectués par lots de 11 à 22 échantillons. Afin d'éviter la confusion d'effet de lots et de fertilité, chaque lot contient une proportion équivalente d'échantillons issus de taureaux fertiles et subfertiles. Pour les trois cohortes présentées ci-dessus, plusieurs éjaculats qui étaient eux-mêmes représentés par plusieurs paillettes ont été réunis afin de constituer un échantillon, de manière à masquer des effets très ponctuels n'affectant qu'un seul éjaculat.

A ces 3 cohortes homogènes s'ajoute une petite cohorte d'éjaculats individuels collectés à des âges variables, composés de 20 taureaux de race Montbéliarde et de fertilité contrastée (4 subfertiles et 16 fertiles). L'équivalent de cette cohorte en Holstein est en cours de constitution. Cette cohorte de

taureaux de race Montbéliarde n'a pas pu être utilisée en intégration de données, car les technologies utilisées pour l'analyse des sncRNA dans cette cohorte et pour la cohorte fertilité décrite ci-dessus sont différentes. Ce n'est pas le cas pour la méthylation de l'ADN ; cette petite cohorte a donc été utilisée comme cohorte indépendante pour évaluer le modèle de prédiction sur des éjaculats individuels (article 1).

Une fois les animaux classifiés, le contenu des paillettes récolté et les échantillons constitués, différentes analyses ont été effectuées. Dans cette thèse ont été analysés la méthylation de l'ADN, les sncRNAs, les paramètres spermatiques et le génotype des animaux. La méthylation de l'ADN, les sncRNAs et les paramètres spermatiques ont été analysés sur les différents échantillons de semence préparés à partir des paillettes décongelées. Les génotypes eux, sont la propriété des entreprises de sélection et ont été établis de manière indépendante par un prestataire de service (LABOGENA), le plus souvent à partir d'un prélèvement de cartilage d'oreille des animaux. Il est important de noter qu'après l'analyse des paramètres spermatiques, les cellules sont culottées, le diluant permettant de conserver la semence est lavé, et les éventuelles cellules somatiques qui contamineraient la semence sont éliminées par un lavage en eau. Cette contamination somatique est rare dans la semence bovine utilisée en IA, et presque indétectable par une observation microscopique. Le lavage en eau est une précaution supplémentaire pour s'assurer que les molécules extraites proviennent exclusivement des spermatozoïdes, qui résistent au choc osmotique contrairement aux cellules somatiques. Nous avons confirmé l'absence de contamination somatique dans l'ADN génomique extrait en examinant le patron de méthylation de gènes importants pour la différenciation germinale (voir figure supplémentaire 2 de l'article 1, Annexe 3).

La partie suivante a pour but de présenter les différents types de données analysées dans ce travail de thèse. Encore une fois, la méthylation de l'ADN fait l'objet d'une présentation plus détaillée car son étude a été plus approfondie au cours de la thèse.

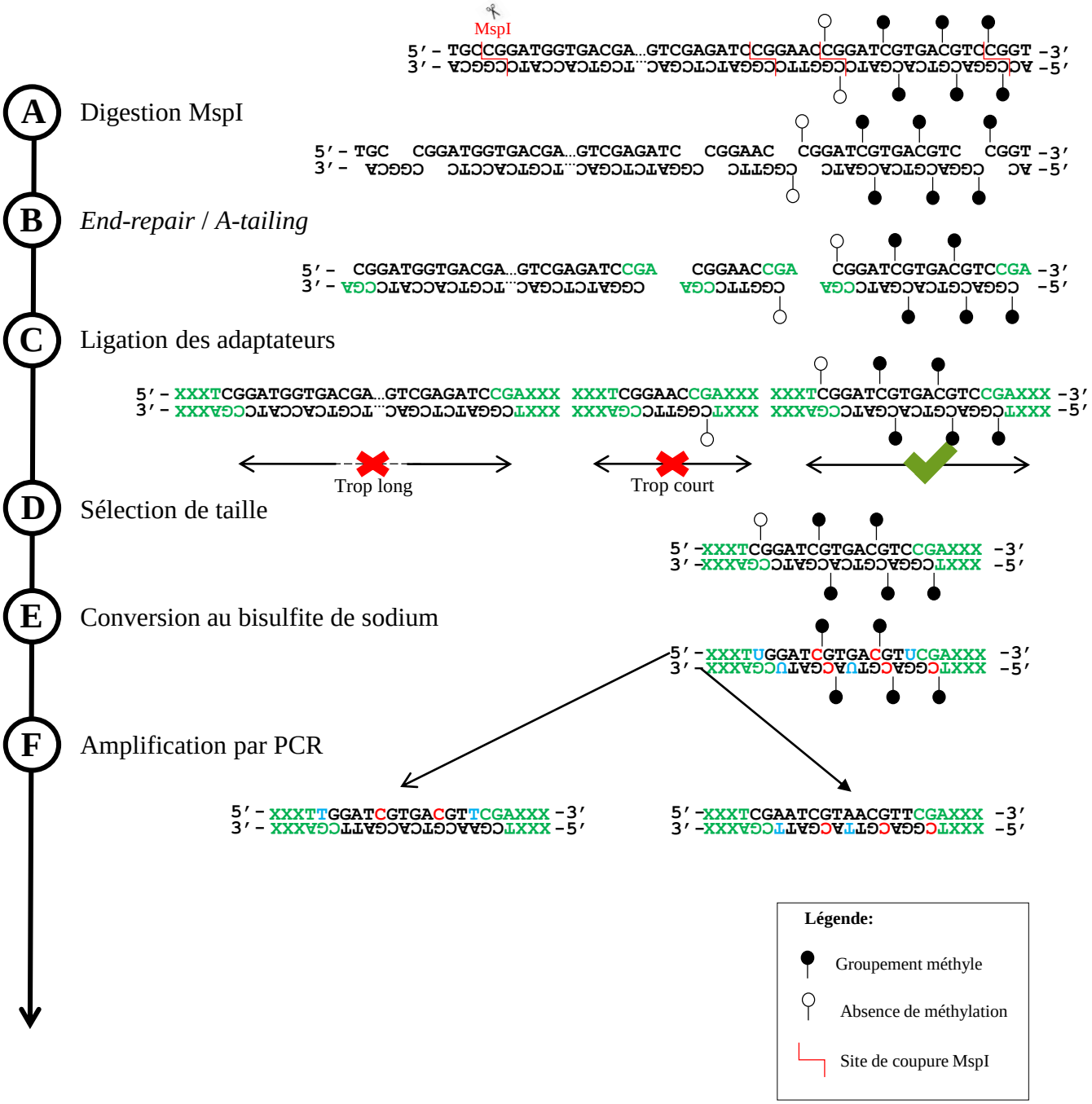


Figure 21 : Les différentes étapes de l'analyse de la méthylation de l'ADN par RRBS. A : L'ADN génomique est dans un premier digéré par l'enzyme MspI qui clive les sites C^CGG, ce qui permet d'avoir un CG en début de séquence. B : Une étape de « End repair » et de «A tailing », permettent de préparer les séquences pour la fixation des adaptateurs Illumina (étape C). D : Une sélection de taille est effectuée permettant de sélectionner les fragments entre 40 et 290 pb. Une fois les fragments sélectionnés, une conversion au bisulfite de sodium est réalisée (étape E) afin de différencier les cytosines méthylées et les cytosines non méthylées. A l'issue de l'amplification par PCR, les cytosines méthylées resteront cytosines et les cytosines non méthylées seront converties en thymine. (Tiré de la thèse de Jean-Philippe Perrier, 2017).

Méthylation de l'ADN

La technique utilisée pour analyser le méthylome spermatique est le RRBS pour « Reduced Representation Bisulfite Sequencing ». Cette méthode est qualifiée de pan-génomique car elle analyse le profil de méthylation de CpG répartis sur tout le génome. Cependant, cette analyse n'est pas exhaustive car elle se concentre principalement sur les régions denses en CpG (Meissner *et al.*, 2005). Le protocole détaillant la mise en œuvre de cette technique au laboratoire se trouve dans la partie « Methods » de l'article 1 se trouvant dans la partie I.II, tandis que les grands principes en sont exposés ci-dessous (Figure 21).

Dans un premier temps, l'ADN génomique extrait des spermatozoïdes est digéré à l'aide de l'enzyme MspI. Cette enzyme clive l'ADN au niveau de sites C^CGG, ce qui permet aux fragments d'avoir au moins un CpG en début de séquence (site potentiel de méthylation de l'ADN). Ces séquences, après ligation avec les adaptateurs Illumina, sont ensuite soumises à une sélection de taille ciblant les fragments génomiques entre 40 et 290 bp. Les séquences clivées commençant forcément par un site CpG, cela implique que la distance maximale entre deux sites CpG sélectionnés est de 290 bp. Cela explique donc pourquoi le RRBS cible les régions génomiques riches en CpG. Le seuil inférieur de 40 pb est établi pour éviter la sélection de trop petits fragments qui poseraient des difficultés d'alignement. A la suite de cette étape, l'objectif est de distinguer les cytosines méthylées des cytosines non méthylées. Pour cela, les fragments d'ADN sont soumis à un traitement au bisulfite de sodium, qui agit de manière différentielle en fonction de la présence ou non d'un groupement méthyle sur les cytosines. En effet, dans les conditions d'application de ce traitement les cytosines méthylées ne sont pas affectées, en revanche les cytosines non méthylées sont converties en uracile. L'ADN est finalement amplifié par PCR afin d'augmenter la concentration et d'introduire les index Illumina permettant de séquencer plusieurs échantillons de manière simultanée puis de réattribuer les séquences produites à chaque échantillon (« démultiplexage »). Ainsi à l'issue de cette amplification les cytosines non méthylées sont converties en thymine et les cytosines méthylées restent sous forme

de cytosine, ce qui permettra de les différencier. Ces séquences sont ensuite séquencées à l'aide d'un appareil HiSeq de marque Illumina, puis alignées sur un génome de référence converti *in silico* au bisulfite à l'aide d'un aligneur spécifique (Bismark (Krueger and Andrews, 2011)). Cet aligneur produit en sortie un fichier listant l'ensemble des CpG couverts par au moins une séquence s'alignant de manière unique sur le génome, les coordonnées chromosomiques du CpG, le nombre total de séquences couvrant le CpG (qu'elles soient porteuses d'un « C » ou d'un « T » au niveau du CpG) et le nombre total de séquences porteuses d'un « C ». Il est important de noter que le résultat de l'alignement de ces séquences converties au bisulfite est complètement confondu avec le résultat d'un polymorphisme C>T affectant le CpG (polymorphismes les plus fréquents). Ce constat motive l'analyse de l'impact direct de sites CpG polymorphes sur les données de méthylation de l'ADN (partie Résultats, Prise en compte des CpG polymorphes).

Les données de méthylation sont particulières car, d'un point de vue biologique, à un site CpG donné, l'information de méthylation est binaire pour les spermatozoïdes qui sont des cellules haploïdes : soit la cytosine est méthylée soit elle ne l'est pas, alors que 3 états de méthylation sont possibles pour les cellules diploïdes (non méthylé sur les 2 allèles, méthylé sur un des 2 allèles, méthylé sur les 2 allèles). Cependant l'ADN génomique a été extrait à partir d'une population cellulaire composée de multiples spermatozoïdes qui peuvent chacun être porteur d'un C méthylé ou non méthylé à chaque CpG. A un CpG donné donc, l'intégration de l'ensemble des états de méthylation au niveau de tous les spermatozoïdes extraits produit un pourcentage de méthylation qui est une variable continue. Ce pourcentage reflète le statut de méthylation d'un CpG dans l'ensemble de la population cellulaire analysée. Il est calculé après l'alignement des séquences sur le génome de référence à partir du fichier de sortie de Bismark, en comptant les séquences contenant un C parmi l'intégralité des séquences couvrant ce CpG. Ainsi, lorsque l'on évoque 80% de méthylation au niveau d'un CpG donné, cela signifie que parmi les séquences couvrant ce CpG 80% présentent un C, ce qui reflète le fait que dans la population de spermatozoïdes analysés, 80% présentent une méthylation au niveau du CpG. Pour assurer une précision minimale dans le calcul du pourcentage de méthylation, les CpG couverts par

moins de 10 séquences ne sont pas pris en compte. Ainsi, en sortie de pipeline pour chaque individu, pour les CpG analysés étant au moins couverts par 10 séquences, une valeur de méthylation exprimée en pourcentage est associée.

Les petits ARN non codants

Au cours de ce travail de thèse, les scnRNA ont été analysés sur les mêmes échantillons que ceux utilisés pour l'analyse de la méthylation de l'ADN. Le protocole d'extraction des ARN totaux et le pipeline d'analyse/annotation ont été développés au cours de la thèse d'Eli Sellem et sont précisément décrits dans Sellem *et al.,* (2020). Ce pipeline a pour objectif, à partir des informations de séquençage, de construire une table de comptage pour chaque fragment ayant été séquencé et correspondant à un sncRNA, ainsi que l'attribution à un type donné de sncRNA.

Dans un premier temps, un filtre est effectué sur la taille des séquences, excluant toutes les séquences inférieures en taille à 17 nt, et sur la qualité des séquences en supprimant les séquences ayant au moins une base avec un Phred score inférieur à 25 (indicateur de la qualité de la séquence). A l'issue de cette étape, toutes les séquences uniques sont comptées permettant *in fine* d'obtenir une table référençant pour chaque échantillon, les séquences retrouvées et leur comptage respectif. Dans le but d'obtenir une caractérisation précise de chaque séquence, elles sont alignées sur différentes bases de données (Ensembl, RFAM, tRNA bovin de GtRNAdb, ncRNA bovin de ENA et piRNA bovin, murin et humain de ENA). Parallèlement à ces étapes, le logiciel mirDeep2 est utilisé dans le but de caractériser spécifiquement les miRNA (Friedländer *et al.*, 2012). Cette partie du pipeline d'analyse a pour objectif à la fois d'annoter les miRNA (miRBase) mais également d'identifier leurs localisations génomiques.

En sortie de pipeline seront donc disponibles pour chaque échantillon, les séquences analysées, leur comptage respectif ainsi qu'une information d'annotation.

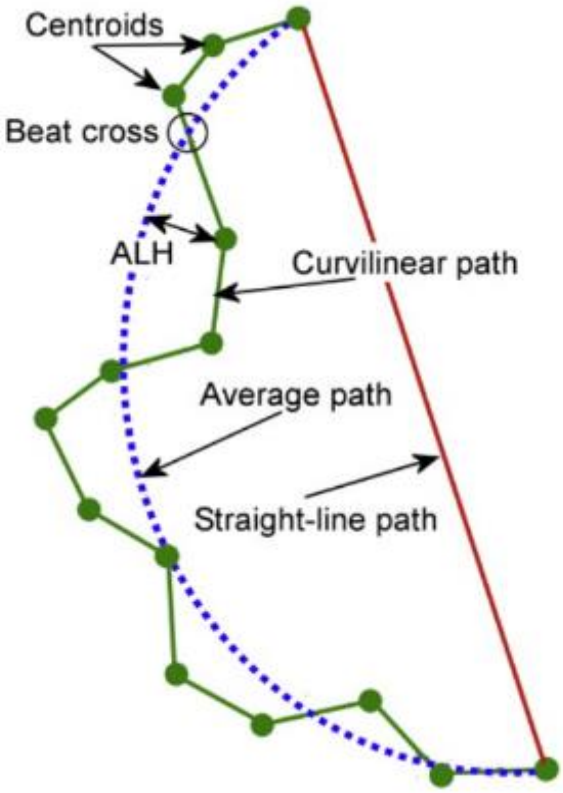


Figure 22 : Les paramètres spermatiques. Schéma représentant les différents éléments mesurés pour caractériser la qualité du mouvement et de la dynamique des spermatozoïdes. Figure inspirée de (tiré de Amann et al. ; 2014)

Les paramètres spermatiques

Au moment de la préparation des échantillons de semence dans le but d'analyser la méthylation et les scnRNA, différents paramètres spermatiques sont analysés une fois les paillettes décongelées et rassemblées et avant le retrait du diluant. Les paramètres spermatiques sont de natures hétérogènes car ils mesurent différentes propriétés, sont exprimés dans différentes unités et sont analysés par CASA (« Computer Assisted Semen Analysis ») et par cytométrie en flux. Les paramètres spermatiques représentent un jeu de données de petite dimension composé de 15 variables.

Il y a 7 variables qui sont mesurées à l'aide du CASA (Figure 22) : Le pourcentage de spermatozoïdes motiles ; le pourcentage de spermatozoïdes progressifs ; « VAP » (la vitesse de trajectoire moyenne) exprimée en µm/s; « VSL » (la vitesse de trajectoire en ligne droite) exprimée en µm/s; « VCL » (la vitesse de la trajectoire curvilinéaire) exprimée en µm/s ; « ALH » (amplitude des mouvements latéraux de la tête) ; « STR » (la rectitude du déplacement). Ces paramètres ont pour objectif de caractériser la dynamique du mouvement des spermatozoïdes qui est un critère important de la qualité de la semence.

Les 8 autres paramètres ont été étudiés par cytométrie en flux et s'intéressent à la viabilité (présence d'iodure de propidium dans la cellule, qui permet d'établir un pourcentage de cellules mortes), à l'activité mitochondriale et à la sensibilité à l'oxydation et sont également des témoins de la qualité du spermatozoïde.

Ces paramètres spermatiques ont été sélectionnés parmi un ensemble plus vaste de mesures pouvant être obtenues sur le même appareillage, car ils semblent être les plus prédictifs de la fertilité sur le terrain (Sellem e*t al.,* 2015).

Les génotypes

Tous les taureaux inclus dans les cohortes du programme SeQuaMol sont les propriétés d'entreprises de sélection. Ainsi ces animaux ont été sélectionnés car porteurs d'une génétique d'intérêt pour des caractères agronomiques. L'information génomique de ces animaux est donc disponible. C'est en général à partir de cartilage d'oreille pendant les premiers mois de vie de ces animaux que le génotypage est réalisé en utilisant la puce de génotypage EuroGMD 50K (Illumina), qui permet d'obtenir l'information de polymorphismes de séquence sur environ 50 000 « Single Polymorphism Nucleotides » (SNP). Les SNP sont des polymorphismes de séquence n'affectant qu'un seul nucléotide et avec deux allèles possibles, un allèle de référence et un allèle alternatif. Les données issues de ce génotypage sont de nature tri-modale et qualitative avec comme notation pour un SNP donné : 0 (homozygote sur l'allèle de référence) ; 1 (hétérozygote) ; 2 (homozygote sur l'allèle alternatif). A l'issue de cette analyse, l'information de génotype des animaux est donc disponible sur environ 50 000 SNP. Cependant, à l'aide de stratégies d'imputation qui permettent d'estimer le génotype de certains SNP non présents sur la puce, il est possible d'augmenter le nombre de positions pour lesquelles une information de génotype est disponible. L'imputation est rendue possible grâce à la propriété du déséquilibre de liaison (le fait que les allèles de 2 loci proches soient co-transmis de manière préférentielle), d'une base de données de référence ainsi que des liens de parenté des animaux de cette cohorte (Logiciel Minimac4). La qualité de l'imputation est hétérogène selon les positions génomiques et dépend par exemple de la distance avec le plus proche SNP génotypé. Ainsi, en utilisant l'imputation il est possible d'étendre le nombre de positions génomiques pour lesquelles l'information génétique est disponible à plusieurs millions de loci. Dans ce travail de thèse, nous nous sommes placés à deux niveaux différents : aux génotypes obtenus sur puce et à la séquence imputée.

Résultats

I Le méthylome spermatique et son utilisation dans la prédiction de fertilité des taureaux.

I.I : Prise en compte et traitement des CpG polymorphes

I.I.I : Contexte

La méthylation de l'ADN et la génétique sont liées l'une à l'autre de différentes manières. Il a en effet été montré que des facteurs génétiques contrôlaient une partie du méthylome spermatique (Orozco *et al.*, 2015; Hannon *et al.*, 2018) ou d'autres types cellulaires. Sur le plan mécanistique, ces contrôles peuvent se faire en *cis*, c'est-à-dire qu'un SNP proche d'un CpG contrôle localement son niveau de méthylation. Il peut également y avoir des contrôles en *trans*, par exemple dans le cas où des gènes de la famille des DNMT sont impactés par des SNP modifiant l'activité enzymatique, ce qui aurait un impact plus global. Un autre type d'interactions peut exister, qui lui est directement dû à la présence d'un polymorphisme de séquence sur un site CpG. En effet, si la cytosine et/ou la guanine du site CpG est remplacée par n'importe quel autre nucléotide, alors la méthylation de l'ADN n'est plus possible. Cela peut avoir de grandes conséquences sur les données analysées, qui reflètent alors une variabilité génétique plutôt qu'épigénétique.

Quantifier ces sites CpG polymorphes et prendre en considération ce biais génétique est important lorsque l'on s'intéresse à la méthylation de l'ADN dans un contexte non isogénique, en particulier si des approches de conversion au bisulfite sont mises en œuvre. Ce biais est actuellement négligé chez le bovin, mais il est généralement traité chez l'Homme par l'élimination des sites CpG potentiellement polymorphes. C'est pour cela que cette problématique est exposée en amont de l'analyse du méthylome spermatique dans le cadre de l'étude de la fertilité mâle.

Trois objectifs ont été poursuivis dans ce premier travail : observer la conséquence d'un polymorphisme de séquence sur la distribution de la méthylation d'un CpG, quantifier l'impact des

polymorphismes de séquence sur les données de RRBS et proposer une stratégie afin de prendre en compte l'information de ces polymorphismes de séquence.

Ce travail a été mené en collaboration avec plusieurs membres de l'équipe G2B de l'unité GABI.

I.I.II : Cohortes utilisées

Cette étude a été réalisée en utilisant les animaux de la cohorte « fertilité » des races Montbéliarde et Holstein. Les données utilisées pour réaliser ce travail sont la méthylation de l'ADN et le génotype imputé à la séquence des animaux. Nous avons également exploité les données des 1000 Génomes, qui ne donnent pas accès aux génotypes des animaux, mais sont plus exhaustives en termes de polymorphismes répertoriés dans la population bovine.

I.I.III : Résultats

<u>Identification de CpG différentiellement méthylés</u>

L'objectif premier de l'analyse du méthylome spermatique était d'identifier des CpG différentiellement méthylés entre les animaux fertiles et subfertiles. Pour cela nous avions conduit une analyse différentielle de méthylation entre des taureaux fertiles et des taureaux subfertiles dans la race Montbéliarde. Les analyses différentielles de méthylation, sont des analyses visant à identifier des cytosines individuelles différentiellement méthylées entre deux groupes (ici fertiles contre subfertiles) Le résultat présenté dans ce paragraphe est décrit plus précisément dans l'article 1 (Partie résultat : « Identification of fertility-related differentially methylated CpGs and regions **»**). A l'issue de cette analyse différentielle, nous avons pu identifier 3252 DMC (Cytosines Différentiellement méthylées) entre groupe de fertilité. En regardant en détail ces DMC, nous nous sommes aperçus qu'une partie non négligeable d'entre elles présentaient un profil de type « génétique » (détaillé dans le paragraphe ci-dessous). En se basant sur les génomes imputés à notre disposition, nous avons pu remarquer que 2352 de ces DMC co-localisaient avec des sites polymorphes chez les taureaux analysés, ce qui représente 72% des DMC analysées. Cette proportion importante nous a alertés sur la nécessité de

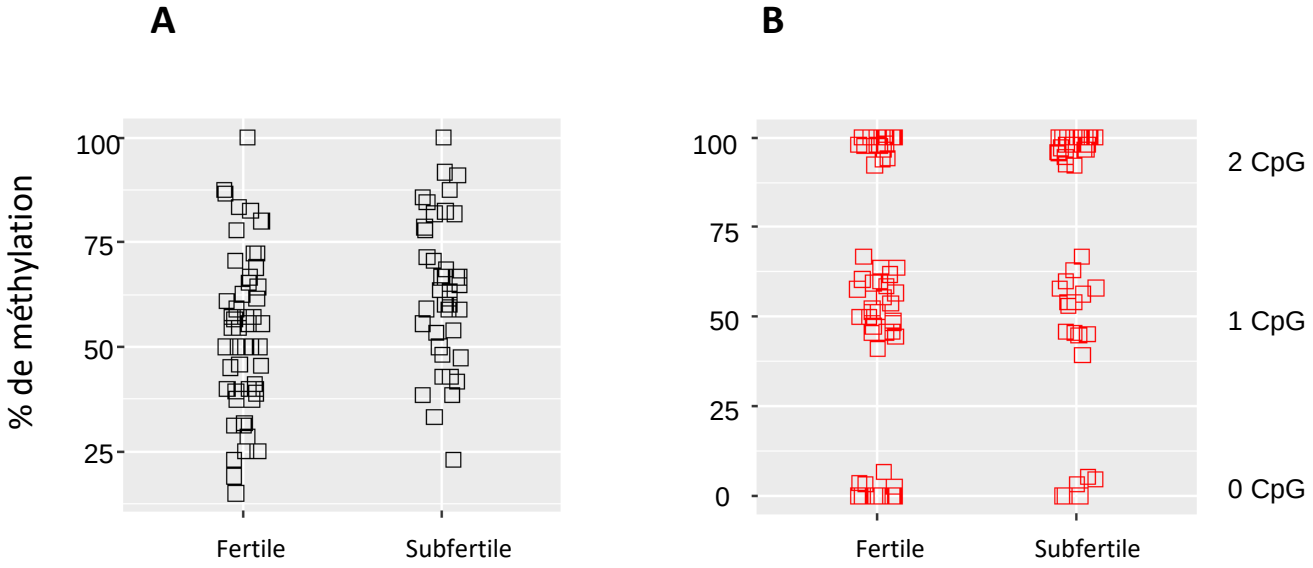


Figure 23 : Impact d'un polymorphisme de séquence sur la distribution de la méthylation de l'ADN. Le profil de méthylation sur un CpG a été analysé au sein d'un site non polymorphe (A) ou polymorphe (B). L'axe des y montre le niveau de méthylation, l'axe des x le statut de fertilité de l'animal, et chaque carré représente le niveau de méthylation pour un animal donné.

trouver des solutions permettant de prendre en compte le polymorphisme de séquence dans l'analyse du méthylome spermatique.

Impact d'un polymorphisme de séquence sur un site CpG

La Figure 23 montre la distribution de la méthylation de l'ADN de deux sites CpG dans la cohorte Montbéliarde. Un seul de ces CpG est polymorphe. Dans la Figure 23A, qui représente le niveau de méthylation du CpG non polymorphe, on observe un profil de méthylation que l'on peut qualifier de continu, ce qui est attendu. Cependant, dans la Figure 23B on observe cette fois-ci un profil de méthylation discret. Ce type de distribution est typique d'un CpG localisé sur un polymorphisme de séquence, avec une distribution des génotypes plutôt équilibrée entre chaque allèle. En effet, les animaux présentant une méthylation de 100% ont obligatoirement un génotype CpG/CpG. Les animaux ayant 50% de méthylation ont très probablement un allèle CpG et un autre allèle non CpG.

En observant ces données, on comprend bien que la présence d'un polymorphisme de séquence au niveau d'un CpG a un effet majeur sur le profil de méthylation des animaux. Il est donc nécessaire de mettre en œuvre une stratégie pour traiter ces CpG polymorphes.

Stratégie proposée

Bien que les polymorphismes aient un impact considérable sur la distribution de méthylation au niveau des CpG avec lesquelles ils co-localisent, un effet épigénétique n'est pas pour autant exclu. En effet, si à un CpG polymorphe donné, il reste un site CpG, alors ce site peut potentiellement accueillir une méthylation ou non. Ainsi en plus de la présence d'un polymorphisme, il peut y avoir une variabilité dans la méthylation de l'ADN également. Etant donnée l'importance des CpG potentiellement polymorphes parmi l'ensemble des CpG analysés par RRBS, il pourrait être intéressant de pouvoir les analyser en fonction du génotype des animaux. En effet, à un CpG polymorphe donné, certains animaux pourraient être homozygotes CpG/CpG (et donc potentiellement présenter un niveau de méthylation variant de 0 à 100%) ou hétérozygotes (niveau de méthylation ≤50%). Il n'y aurait que pour les animaux homozygotes non CpG que l'analyse de la méthylation ne présenterait aucun intérêt.

A

Base concernée	Génotype base	Génotype CpG	Nombre d'allèle CpG (γi)
C	C/C	CpG/CpG	2
C	C/D	CpG/DpG	1
C	D/D	DpG/DpG	0
G	G/G	CpG/CpG	2
G	G/H	CpG/CpH	1
G	H/H	CpH/CpH	0

B

Génotype base 1	Génotype base 2	Génotype CpG	Nombre d'allèle CpG (γi)
C/C	G/G	CpG/CpG	2
C/C	G/H	CpG/CpH	1
C/C	H/H	CpH/CpH	0
C/D	G/G	CpG/DpG	1
C/D	G/H	CpG/DpH ou CpH/DpG	1 ou 0
C/D	H/H	CpH/DpH	0
D/D	G/G	DpG/DpG	0
D/D	G/H	DpG/DpH	0
D/D	H/H	DpH/DpH	0

Figure 24 : Nombre d'allèles CpG au sein d'un CpG polymorphe en fonction du génotype des animaux. A : Table permettant de déduire le nombre d'allèles CpG (donc d'allèles méthylables) dans le cadre où une seule base du site CpG est polymorphe (la cytosine ou la guanine). B : Table permettant de déduire le nombre d'allèles CpG dans le cadre où la cytosine et la guanine sont toutes deux impactées par un polymorphisme de séquence. La lettre H signifie : « tout sauf G » et la lettre D signifie « tout sauf C ».

La difficulté revient à séparer l'information issue du polymorphisme de celle issue de la méthylation de l'ADN.

Le signal de méthylation mesuré à un CpG polymorphe donné peut être décomposé en un effet génétique (lié au nombre d'allèles CpG) et un effet épigénétique (méthylation des allèles CpG).

On peut formuler cette relation de la façon suivante :

$$y_i = \mu + \gamma_i + \varepsilon_i$$

y_i : vecteur des observations (niveau de méthylation d'un individu au CpG analysé)

μ : moyenne de méthylation des individus au CpG analysé

γ_i : vecteur des effets fixes, permettant de répertorier le nombre d'allèles CpG pour chaque animal

ε_i : méthylation résiduelle du site CpG c'est-à-dire le niveau de méthylation restant après avoir supprimé l'influence du SNP (effet épigénétique).

On peut donc en déduire :

$$y_i^* = y_i - (\mu + \gamma_i) = \varepsilon_i$$

Avec y_i* représentant une valeur de méthylation corrigée de l'influence du polymorphisme de séquence, c'est-à-dire la résiduelle du modèle. Ainsi, la connaissance du génotype d'un animal donné à un CpG donné, en renseignant sur le nombre d'allèles CpG, permet théoriquement d'accéder à la méthylation résiduelle. Ce modèle doit être appliqué à l'ensemble des CpG polymorphes analysés

<u>Déduction du nombre d'allèles CpG à partir des séquences imputées</u>

Dans le but de connaître le nombre d'allèles CpG potentiellement méthylables pour un animal au niveau d'un CpG polymorphe, on peut considérer deux cas de figure. Le premier est le cas où le CpG n'est l'objet que d'un SNP touchant le C ou le G. Le deuxième cas, plus rare, est le cas où la cytosine et la guanine sont toutes deux polymorphes. Le détail de la déduction du nombre de site méthylable dans ces deux cas de figures en fonction des génotypes des animaux est présenté en Figure 24.

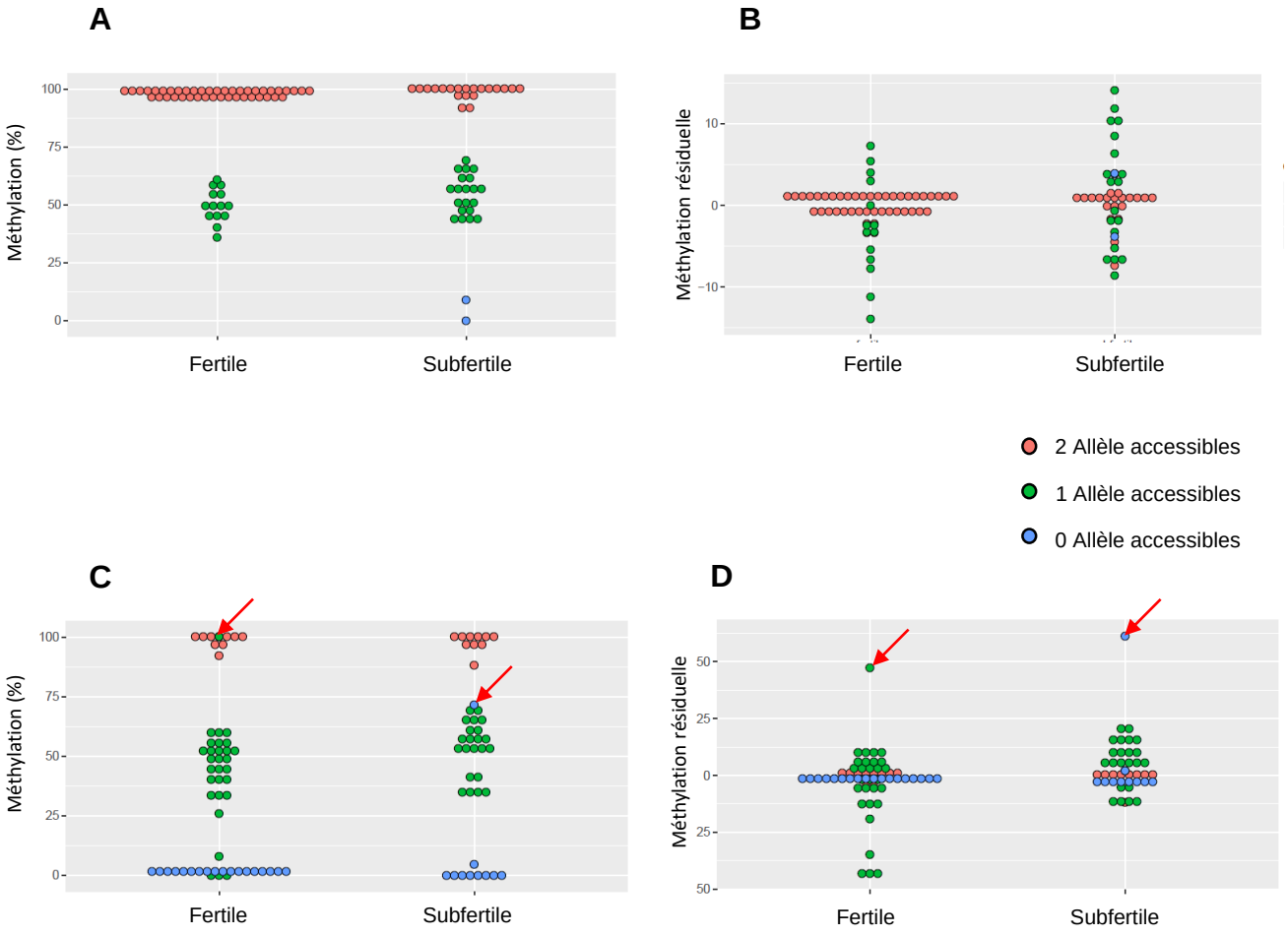


Figure 25 : Exemple de profils de méthylation avant et après correction par les génotypes. A et B : Niveau de méthylation au sein du même CpG polymorphe. Chaque point représente un individu et est colorié en fonction du nombre d'allèles CpG. En A il s'agit des données brutes, et en B des données corrigées du polymorphisme de séquence. En C il s'agit d'un autre CpG polymorphe avec des incohérences entre le nombre d'allèles CpG et le niveau de méthylation mesuré. Les deux flèches rouges indiquent des situations théoriquement impossibles. D : Il s'agit du CpG présent en C après correction. Les flèches rouges indiquent les points mis en évidence en C.

Ainsi à l'issue de cette étape il est possible pour chaque CpG polymorphe de connaitre le nombre d'allèles portant un site CpG pour chaque individu et donc d'extraire la méthylation résiduelle.

Analyse des données de méthylation après correction par le génotype

A la suite de ce travail nous avons analysé les données corrigées. Afin de montrer un exemple, le niveau de méthylation d'un CpG polymorphe avant et après correction est présenté dans la Figure 25AB. En 25A, on observe bien les mêmes conséquences que présentées plus haut dans la Figure 23B, où la source de variabilité principale a pour origine le polymorphisme de séquence. On remarque qu'après correction, comme c'est attendu, la variabilité n'est plus influencée par le polymorphisme de séquence, mais par la variabilité existant au sein de chaque classe génotypique. Cela montre que par cette méthode nous avons bien extrait l'information de méthylation résiduelle, en enlevant l'impact du polymorphisme sur les données. Les données corrigées, étant extraites de la résiduelle du modèle, sont centrées.

Cependant, en explorant les données en détail, nous avons pu mettre en évidence certains évènements théoriquement impossibles et cela est illustré dans la figure 25C. En effet, au sein d'un CpG polymorphe on observe bien une distribution discrète du profil de méthylation avant correction. Cependant, on peut observer un « point vert » correspondant à un individu n'ayant qu'un seul allèle CpG, avec un niveau de méthylation proche de 100%. De plus, on observe également un « point bleu », correspondant à un animal n'ayant aucun allèle CpG avec un niveau de méthylation aux alentours de 50%. Dans les deux cas de figures, cela est théoriquement impossible. Cette incohérence entre le profil de méthylation et le génotype peut avoir des conséquences majeures après correction (Figure 25D).

Les incohérences ainsi pointées pourraient remettre en cause la démarche de correction mise en œuvre. A ce stade, il a été nécessaire de quantifier ces incohérences afin de savoir si elles concernaient une partie importante de CpG polymorphes ou non (auquel cas la stratégie de correction aurait éventuellement pu être envisagée).

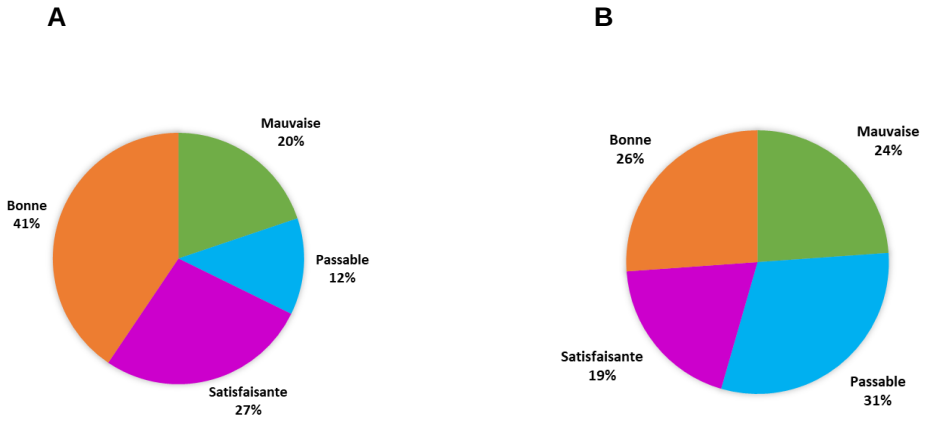


Figure 26 :Les incohérences entre niveau de méthylation et génotype sont majoritaires. Quantification des CpG polymorphes présentant une incohérence entre niveau de méthylation et génotypes dans la race Holstein (A) et la race Montbéliarde (B). Les CpG ont été classifiés en 4 classes différentes : Bonne : Aucune incohérence entre le génotype et le profil de méthylation ; Satisfaisante : Globalement une bonne cohérence, seuls quelques individus incohérents ; Passable : Davantage d'individus incohérents, dans la limite de 50% d'individus cohérents ; Mauvaise : Plus de la moitié d'individus incohérents Déplacer la descrition des différentes catégories ici

Quantification des incohérences entre profils de méthylation et génotypes

Ces incohérences ont été évaluées en races Montbéliarde et Holstein. Toute la population des DMC polymorphes n'a pas été examinée : en race Montbéliarde ce travail a été réalisé sur 134 DMC polymorphes et en Holstein sur 217 DMC polymorphes. Les CpG polymorphes ont été classés en 4 classes en fonction de la cohérence entre génotypes et profils de méthylation :

Un seul opérateur s'est chargé de faire la classification afin de limiter les biais. Les résultats de cette analyse sont présentés dans la figure 26AB.

Dans les deux races, on observe que le nombre de CpG pour lesquels profils de méthylation et génotypes sont incohérents représentent plus de la moitié des CpG polymorphes analysés. De plus on observe une différence entre races, car les CpG de classe « Bonne » représentent 41% des CpG en Holstein mais seulement 26% dans la race Montbéliarde.

Origine des incohérences entre profils de méthylation et génotypes

Ces incohérences peuvent soit provenir d'erreurs dans l'estimation du pourcentage de méthylation, soit provenir d'erreurs dans les génotypes des animaux. Comme expliqué précédemment l'estimation du pourcentage de méthylation est basée sur le comptage de séquences portant soit un C soit un T au niveau d'un CpG donné. La précision de cette mesure est donc soumise à la profondeur de séquençage (nombre de séquences alignées sur un CpG donné). Nous avons pu éliminer d'autres facteurs, comme par exemple le taux de conversion au bisulfite, qui s'est avéré être d'un excellent niveau pour tous les échantillons (Table 1 de l'article 1 (Montbéliard) et Annexe 5 (Holstein)). Il est possible que les CpG incohérents soient ceux pour lesquels la profondeur est la plus faible, donnant lieu à des imprécisions dans la valeur de méthylation mesurée. L'autre possibilité est une erreur dans les génotypes des animaux, ce qui est d'autant plus plausible que pour une grande majorité des CpG, les génotypes ont été obtenus par imputation et non par génotypage sur la puce. Si à l'échelle d'un génome et d'une population l'imputation est fiable, elle peut constituer une source d'erreurs à l'échelle d'un locus et d'un animal.

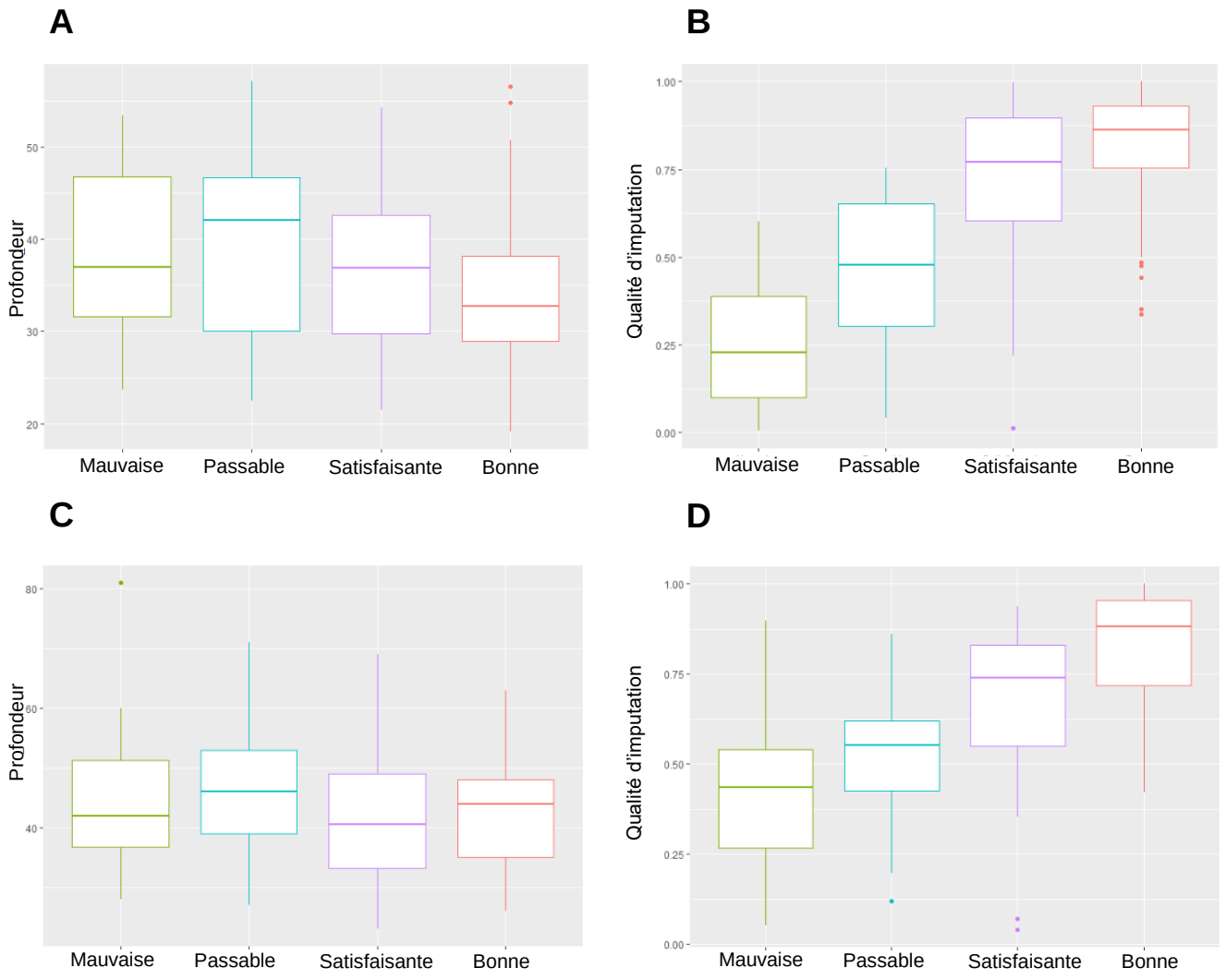


Figure 27 : La qualité d'imputation est le facteur expliquant une partie des incohérences entre niveaux de méthylation et génotypes. La profondeur de séquençage ainsi que la qualité d'imputation ont été représentées en fonction des classes attribuées aux CpG polymorphes dans la race Holstein (A et B) et Montbéliard (C et D). Si aucune tendance ne se dégage pour la profondeur de séquençage, on observe que les incohérences diminuent avec la qualité d'imputation.

Afin de déterminer s'il y a une association entre ces 2 sources d'erreur et la cohérence entre profils de méthylation et génotypes, nous avons analysé les variations de profondeur et de qualité d'imputation dans les 4 classes de CpG polymorphes définies précédemment (Figure 27).

Dans les deux races, les conclusions sont les mêmes : on n'observe pas d'association entre la profondeur et la classe de cohérence du CpG, mais une association claire entre la qualité de l'imputation et la classe de cohérence du CpG. Ces résultats démontrent donc qu'une partie des facteurs contribuant à l'incohérence entre le profil de méthylation et le génotype a pour origine la qualité de l'imputation.

Ce résultat permet également de donner du sens au fait que l'on trouve plus de CpG polymorphes classés avec une cohérence « Bonne » en Holstein qu'en Montbéliarde. En effet, la population permettant l'imputation en race Holstein est plus large qu'en race Montbéliarde, ainsi la qualité de l'imputation en Holstein est meilleure.

Conclusion

Pour conclure sur ce travail, nous avons donc proposé une méthode permettant d'extraire l'information de méthylation résiduelle (non liée à de la variabilité génétique) au sein de sites CpG polymorphes. Nous avons mis en évidence des problèmes de cohérence entre le génotype des animaux et le profil de méthylation associé. Ces incohérences ont en partie pour origine la qualité de l'imputation des génotypes des animaux. Ces incohérences touchent une majorité de CpG polymorphes, et faussent complètement la correction des données de méthylation. Cette stratégie de correction ne peut donc pas être envisagée pour séparer l'information génétique de l'information de méthylation. Il a donc été décidé dans ce travail de thèse de supprimer l'intégralité des CpG co-localisant avec des polymorphismes référencés dans l'espèce bovine. Pour cela nous nous sommes basés sur la base de données des 1000 Génomes Bovins, référençant un grand nombre des polymorphismes de séquence ayant été mis en évidence dans l'espèce bovine. Nous avons croisé l'intégralité des CpG couverts par au moins une séquence dans nos analyses, avec les données des

1000 Génomes Bovins, et éliminé 22% de CpG potentiellement polymorphes (sur respectivement 4 502 501 et 4 090 371 CpG couverts dans les races Montbéliarde et Holstein, 986 748 et 881 670 CpG).

Néanmoins, le travail réalisé dans cette partie n'a pas été inutile. Nous avons en effet pu montrer qu'il était possible d'extraire une partie de l'information de méthylation contenue dans un CpG polymorphe à partir des connaissances du génotype. Cette procédure de correction pourrait être intéressante dans des situations où le génotype est évalué de manière précise à chaque locus (par exemple, par re-séquençage du génome complet).

I.II : Analyse du méthylome spermatique en relation avec la fertilité mâle

I.II.I : Contexte :

Dans l'espèce bovine, la fertilité mâle est un phénotype important tant sur le plan économique qu'organisationnel, du fait de l'utilisation de l'IA à grande échelle pour la reproduction. Des études ayant pour but d'analyser la génétique et les paramètres spermatiques ont été réalisées afin de prédire la fertilité. Comme expliqué dans l'introduction, l'héritabilité de la fertilité mâle est faible ce qui fait qu'il n'existe pas d'évaluation génomique de la fertilité mâle appliquée en routine aujourd'hui. D'un autre coté la méthylation de l'ADN est importante pour la fertilité mâle car impliquée dans la différenciation des cellules germinales mâles, dans la spermatogénèse et le développement embryonnaire. La reprogrammation à laquelle elle est soumise au cours de la différenciation des cellules germinales la rend également sensible à l'exposition de ces cellules à un environnement délétère, qui pourrait ainsi affecter la fertilité à l'âge adulte et influencer le phénotype à long terme de la descendance.

Ainsi, l'objectif dans cette partie est d'analyser le méthylome spermatique de taureaux fertiles et de taureaux subfertiles afin d'identifier des marques de méthylation différentielles en fonction de la fertilité. Ces marques ont été analysées afin de mettre en évidence les régions génomiques et les gènes touchés dans le but de les relier à des fonctions biologiques. De plus un modèle de prédiction a été

construit à partir de ces marques différentielles. Ce travail a été réalisé dans deux races différentes : la race Montbéliarde et la race Holstein. Les résultats obtenus en race Montbéliarde ont fait l'objet d'une publication acceptée (Costes *et al.*, 2022). L'article est présenté dans le chapitre suivant et un résumé des principaux résultats est présenté ci-dessous. Les résultats en race Holstein n'ont pas donné lieu à une publication et sont présentés à la suite de l'article en race Montbéliarde.

Résumé de l'article en race Montbéliarde :

Dans la race Montbéliarde, 100 taureaux fertiles et subfertiles de la cohorte « fertilité » ont été analysés. Le méthylome spermatique a été obtenu par RRBS, puis soumis à une analyse différentielle de méthylation.

Cette analyse a mis en évidence 490 DMC entre conditions de fertilité sur un total de 1 548 563 CpG analysés avec une couverture du dispositif et une profondeur de séquençage suffisantes (« background », obtenu après filtre des variants répertoriés dans les 1000 Génomes). Nous avons tout d'abord montré que les DMC étaient présentes sur tous les chromosomes du génome et qu'il n'y avait pas d'enrichissement spatial.

En analysant les éléments génomiques impactés par ces DMC, nous avons pu mettre en évidence des différences en termes de proportion entre les DMC et le background. Certaines régions sont surreprésentées parmi les DMC comme par exemple les régions introniques et intergéniques tandis que d'autres sont sous-représentées, comme par exemple les promoteurs des gènes. Nous avons trouvé un enrichissement important de certaines séquences répétées comme les LINE et les LTR dans les DMC hypométhylées chez les animaux subfertiles.

Nous avons pu identifier 139 gènes différentiellement méthylés (c'est-à-dire contenant au moins une DMC), et à l'aide d'une bibliographie systématique nous avons pointé que 48 d'entre eux avaient été décrits en lien avec le développement embryonnaire et la fonction spermatique.

Enfin, sur la base de 107 DMC sans données manquantes (donc couvertes par un minimum de 10 séquences chez chacun des 100 taureaux) nous avons construit des modèles de prédiction dans le but de classifier un animal comme étant fertile ou subfertile à l'aide de son profil de méthylation. Il a ainsi été possible de prédire la fertilité des animaux avec une précision de 72%, et ce pour deux cohortes différentes (la cohorte « fertilité », ainsi que la petite cohorte indépendante de 20 taureaux décrite dans la partie « Cohorte d'analyse »).

Les Figures et table supplémentaire de cet article sont disponibles en Annexe 3.

I.II.II : Article 1

Costes et al. Clinical Epigenetics (2022) 14:54
https://doi.org/10.1186/s13148-022-01275-x

Clinical Epigenetics

RESEARCH Open Access

Predicting male fertility from the sperm methylome: application to 120 bulls with hundreds of artificial insemination records

Valentin Costes[1,2,3,4], Aurélie Chaulot-Talmon[1,2], Eli Sellem[1,2,3], Jean-Philippe Perrier[1,2], Anne Aubert-Frambourg[1,2], Luc Jouneau[1,2], Charline Pontlevoy[1,2], Chris Hozé[3,4], Sébastien Fritz[3,4], Mekki Boussaha[4], Chrystelle Le Danvic[3], Marie-Pierre Sanchez[4], Didier Boichard[4], Laurent Schibler[3], Hélène Jammes[1,2], Florence Jaffrézic[4] and Hélène Kiefer[1,2*]

Abstract

Background: Conflicting results regarding alterations to sperm DNA methylation in cases of spermatogenesis defects, male infertility and poor developmental outcomes have been reported in humans. Bulls used for artificial insemination represent a relevant model in this field, as the broad dissemination of bull semen considerably alleviates confounding factors and enables the precise assessment of male fertility. This study was therefore designed to assess the potential for sperm DNA methylation to predict bull fertility.

Results: A unique collection of 100 sperm samples was constituted by pooling 2–5 ejaculates per bull from 100 Montbéliarde bulls of comparable ages, assessed as fertile ($n=57$) or subfertile ($n=43$) based on non-return rates 56 days after insemination. The DNA methylation profiles of these samples were obtained using reduced representation bisulfite sequencing. After excluding putative sequence polymorphisms, 490 fertility-related differentially methylated cytosines (DMCs) were identified, most of which were hypermethylated in subfertile bulls. Interestingly, 46 genes targeted by DMCs are involved in embryonic and fetal development, sperm function and maturation, or have been related to fertility in genome-wide association studies; five of these were further analyzed by pyrosequencing. In order to evaluate the prognostic value of fertility-related DMCs, the sperm samples were split between training ($n=67$) and testing ($n=33$) sets. Using a Random Forest approach, a predictive model was built from the methylation values obtained on the training set. The predictive accuracy of this model was 72% on the testing set and 72% on individual ejaculates collected from an independent cohort of 20 bulls.

Conclusion: This study, conducted on the largest set of bull sperm samples so far examined in epigenetic analyses, demonstrated that the sperm methylome is a valuable source of male fertility biomarkers. The next challenge is to combine these results with other data on the same sperm samples in order to improve the quality of the model and better understand the interplay between DNA methylation and other molecular features in the regulation of fertility. This research may have potential applications in human medicine, where infertility affects the interaction between a male and a female, thus making it difficult to isolate the male factor.

Keywords: Male fertility, DNA methylation, Predictive model, Sperm, Cattle

*Correspondence: helene.kiefer@inrae.fr
[1] INRAE, BREED, Université Paris-Saclay, UVSQ, 78350 Jouy-en-Josas, France
Full list of author information is available at the end of the article

© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

Understanding the causes of infertility and subfertility is a challenge in human medicine as these conditions affect approximately 15% of couples. Infertility is defined as an inability to conceive a child after 12 months of unprotected intercourse. In 30% of cases, male factors are recognized as the primary cause of infertility because anatomical, hormonal or known genetic anomalies have been clearly diagnosed but no female factors have been identified [1–3]. However, in 15–30% of cases, male infertility is described as idiopathic [1, 4], meaning that the aforementioned causes have been excluded. Altered semen parameters are commonly observed in cases of idiopathic male infertility, suggesting that unidentified factors that impair spermatogenesis might have compromised fertility [4]. Idiopathic infertility also affects normozoospermic patients, thus making it more difficult to determine the causes and highlighting the need for additional analyses of sperm in order to deepen our understanding of fertility [5].

Mature spermatozoa represent the ultimate form of male germ cell differentiation, which is a tightly regulated process that starts in utero and is achieved in adulthood during spermatogenesis. Sperm cells are meant to survive outside the organism, fertilize an oocyte and contribute to a new individual; and although they are transcriptionally inactive, they bear a remarkable epigenome in line with their high degree of specialization and unique nature [5]. Thus DNA methylation, chromatin, as well as non-coding RNAs, all display unique features inherited from sperm differentiation that are also essential for embryonic development [6, 7]. It has been postulated that alterations to this sperm-specific epigenome due to lifestyle factors, advanced age, medical conditions or environmental exposure may provide an explanation for male infertility or subfertility [8–13]. Of all epigenetic mechanisms, DNA methylation has been a particular focus for studies on male fertility. Manipulations of DNA methylation in pharmacological or genetic models have indeed demonstrated the essential role it plays in male germ cell differentiation and fertility [14–16]. Alterations to the sperm methylome caused by advanced paternal age or exposure to harmful conditions have also been reported to interfere with male fertility, pregnancy outcomes and the health of the next generation [17–21]. Furthermore, from a technical point of view, the starting point for all DNA methylation assays is the preparation of high quality genomic DNA, which is relatively straightforward when compared to other types of epigenetic assays. Standardized platforms such as the Illumina EPIC assay or earlier versions have thus been developed in humans, opening the way to the analysis of large cohorts at a genome scale [22].

Numerous studies in human cohorts have identified specific DNA methylation profiles associated with altered semen parameters [23–25], embryo quality after in vitro fertilization [26], and infertility or subfertility in normozoospermic patients [27, 28]. However, these studies did not reach a consensus regarding a DNA methylation signature for subfertility or infertility [29], with the exception of specific alterations targeting imprinting genes [30], although this view has also been disputed recently [31]. These inconsistent results may have technical or biological origins, such as the diversity of the methods used to process semen and obtain DNA methylation data or the heterogeneity of phenotypes associated with infertility [29]. They may also be related to confounding factors that are inherent to studies on human fertility which make it very difficult to isolate male factors (one male partner, one female partner, both living in the same environment) [32].

Unlike the situation in humans, the widespread use of artificial insemination (AI) in dairy cattle and the dissemination of bull semen to many herds markedly reduces these limitations; the bull is therefore an excellent animal model to investigate the etiology of male subfertility. Indeed, each bull is usually mated with hundreds of cows maintained in different herds, and the outcome of each insemination is recorded. Bull fertility can therefore be measured accurately in the field and corrected for many confounding factors. Moreover, bull fertility is an important economic trait. Unsuccessful AI can indeed give rise to direct costs, extended calving intervals and increased culling rates of the inseminated cows. Considerable efforts have therefore been made in an attempt to predict bull fertility and to understand the causes of subfertility based on genotypes, semen functional parameters or sperm molecular features, including epigenetic modifications [33, 34]. Some studies on alterations to the sperm methylome in bulls belonging to different fertility classes have recently been published [35–40]. Although they highlighted genes potentially important to fertility and were informative from a biological point of view, all these studies involved small numbers of individuals. These studies also investigated a variety of breeds, considered different phenotypes for fertility, and applied a range of technologies to obtain genome-wide DNA methylation profiles, resulting in marked variations in terms of the magnitude of the DNA methylation changes and the nature of impacted genes. As a result, the link between sperm DNA methylation and male fertility remains poorly understood in cattle, despite all the benefits of the bull model.

In order to contribute knowledge in this field, we investigated the nucleotide-level resolution, genome-wide DNA methylation profiles of semen samples collected

from a total of 120 carefully selected Montbéliarde bulls with contrasting fertility levels. We report here on the biological characterization of the loci that were differentially methylated between fertile and subfertile bulls, and on the potential for using these loci to build models predictive of fertility status.

Results

Experimental design and overall strategy

In order to clarify the relationships between sperm DNA methylation and male fertility and to use variations in the methylome to establish a predictive model, 100 semen samples were obtained from marketed Montbéliarde bulls categorized as fertile ($n=57$) or subfertile ($n=43$) according to hundreds of AI records (main cohort, Fig. 1A, B). The bulls were selected based on non-return rates at 56 days (NRR 56), i.e. the proportion of cows that were not re-bred within 56 days of an insemination and could therefore be considered as pregnant. Because these NRR 56 scores were corrected from confounding factors, they fully characterize the bulls and are hardly affected by other sources of variation. The distribution of corrected NRR 56 among all marketed Montbéliarde bulls was normal and quite narrow, and considering its large size and inclusion criteria, the main cohort reflected the most contrasting differences in fertility that could be investigated within this distribution (Fig. 1A). Although the differences in corrected NRR 56 were relatively limited between fertile and subfertile bulls, they were highly significant (Fig. 1B), demonstrating that fertile and subfertile bulls represented two distinct fertility classes.

Several ejaculates representative of the overall fertility of each bull were pooled in order to minimize environmentally or physiologically driven variations that might affect individual ejaculates. To prevent the confounding effect of age on the sperm methylome [41, 42], all ejaculates were collected from animals between 17 and 19 months of age (Additional file 2: Table S1). An assessment of semen functional parameters revealed no statistically significant differences between the fertile and subfertile samples (Additional file 3: Table S2 and Fig. 1C), suggesting that subfertility is not related to defective spermatogenesis. The genome-wide DNA methylation profile of these 100 semen samples was investigated using reduced representation bisulfite sequencing (RRBS) which enabled the identification of fertility-related differentially methylated CpGs (DMCs), which were then annotated and used for functional enrichment analyses and technical validation (Fig. 1D). A Random Forest approach was then applied to the DNA methylation percentages at these DMCs, the aim being to create and validate a model that could predict the fertility class of each bull (Fig. 1E).

The main cohort used to identify DMCs and to build the predictive model was also used to evaluate the model (see "Methods"), which might have led to an overestimation of the model's performance. To overcome this source of bias and assess the potential of the model for field application, an independent and less controlled cohort that did not contribute to identifying the DMCs was also investigated. This independent cohort comprised 20 individual ejaculates collected from 20 marketed Montbéliarde bulls of contrasting fertility (16 fertile and 4 subfertile) and various ages that had not been included in the main cohort (Fig. 1F, G). Like the main cohort, semen functional parameters did not differ significantly between the fertile and subfertile samples (Fig. 1H; Additional file 3: Table S2). After RRBS analysis, the DNA methylation values at the genomic positions previously identified in the main cohort as DMCs were extracted (Fig. 1D). The predictive model built on the main cohort was then applied to these values in order to classify the samples as fertile or subfertile. Model quality indicators were calculated for both cohorts, based on the consistency between predicted and actual fertility classes (Fig. 1E).

(See figure on next page.)

Fig. 1 Experimental design and overall strategy. Fertile and subfertile bulls are shown in red and blue, respectively. **A** distribution of the corrected non-return rate at 56 days post-insemination (NRR 56) for all the Montbéliarde bulls born between 2011 and 2014 (519 bulls in total, in grey) and for fertile ($n=57$) and subfertile bulls ($n=43$) included in the main cohort. **B** The main cohort comprised 100 bulls with contrasting fertility, based on the differences in corrected NRR 56 (Wilcoxon test, $p<0.05$). **C** for each bull, straws representing several ejaculates prepared for AI were thawed and pooled. Semen functional parameters corrected for batch preparation effects (see "Methods") were assessed on these pools and no significant difference in mitochondrial status, motility or viability could be detected between fertile and subfertile bulls (Wilcoxon test, $p \geq 0.05$). **D** DNA methylation was analyzed on these samples by RRBS, enabling the selection of a subset of CpGs at which DNA methylation could be measured with sufficient precision (CpGs10). These CpGs10 were then subjected to differential analysis and fertility-related DMCs were identified. **E** Using the methylation values at DMCs, a predictive model for fertility was constructed and validated. **F** Distribution of the corrected NRR 56 for all the Montbéliarde bulls born between 2009 and 2012 (485 bulls in total, in grey) and for fertile ($n=16$) and subfertile bulls ($n=4$) included in the independent cohort. **G** The independent cohort included 20 bulls with contrasting fertility based on the differences in corrected NRR 56 (Wilcoxon test, $p<0.05$), and was only used to evaluate the potential for field application of the predictive model built on the main cohort. **H** Semen functional analysis and RRBS were performed on one ejaculate per bull. The model previously built on the main cohort was applied to the methylation values obtained in the independent cohort at CpGs10 identified as DMCs using the main cohort. Model quality indicators assessing the consistency between the actual and predicted fertility were then calculated for both the main and independent cohorts

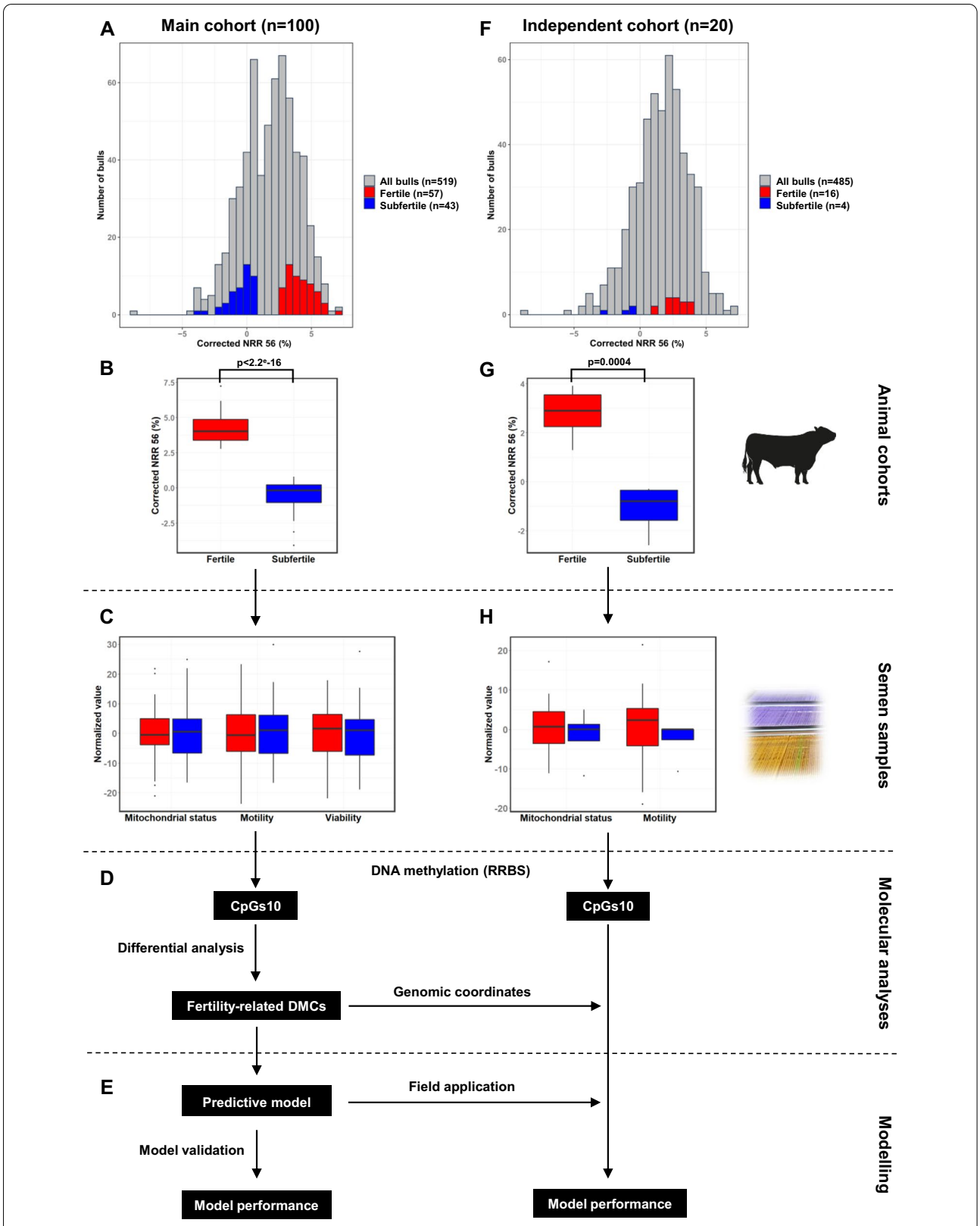


Fig. 1 (See legend on previous page.)

Reduced representation bisulfite sequencing of 120 semen samples

The sequencing parameters, alignment statistics and overall DNA methylation values for both cohorts were analyzed and compared between fertile and subfertile bulls (Table 1; the asterisks indicate significant differences between fertile and subfertile bulls). Sequencing generated an average of 33.6 million read pairs with an average quality (Phred) score of 38.4, meaning that more than 91% of the sequenced bases had high quality scores. The average bisulfite conversion rate reached 99%. The reads were then aligned on the bovine reference genome. Unique mapping efficiency (34.3% on average) was low but consistent with previous RRBS studies in the bovine species [40, 43, 44]. On average, these uniquely mapped reads aligned on 3.2 million CpGs (out of 28 million CpGs in the genome), of which 60.7% on average were covered by at least 10 reads (CpGs10) and were retained for further analysis. This low unique mapping efficiency was due to the high percentage of reads that aligned at multiple locations on the genome (multimapped reads); this feature was attributed to the large number of repetitive elements targeted by RRBS in the bovine species [43]. Because of their repetitive nature in the genome context, repeats are supposed essentially to be covered by multimapped reads that are filtered-out during bioinformatics analysis. This process results in a loss of information regarding these sequences, which are of functional importance and subject to changes that affect DNA methylation in the sperm of infertile men [27, 45]. In order to obtain additional information on the methylation status of repeats in fertile and subfertile bulls, the reads were also aligned on an artificial genome containing one specimen of each repeat, as defined in Repbase [46]. Although the reads aligned with only 1106 CpGs on average, unique mapping efficiency on this Repbase genome reached an average of 20.7%, translating the stacking of a huge amount of reads at each CpG (an average of 38,341 reads per CpG). There were no statistically significant differences regarding sequencing parameters and alignment statistics on the two genomes when comparing fertile and subfertile bulls from either cohort (Table 1). This eliminated the possibility of technical

Table 1 Characterization, mapping efficiency on the bovine reference genome (ARS-UCD2.1) and on a Repbase artificial genome, coverage and average methylation in RRBS libraries

	Main cohort		Independent cohort	
	Fertile ($n=57$)	Subfertile ($n=43$)	Fertile ($n=16$)	Subfertile ($n=4$)
Sequencing parameters				
Number of read pairs (million)	32.9 ± 5.6	33.9 ± 5.4	35.5 ± 5.6	33.9 ± 2.5
Average quality score (Phred score)	38.2 ± 0.3	38.4 ± 0.3	39.0 ± 0.3	38.9 ± 0.4
Bisulfite conversion rate (%)	99.0 ± 0.4	99.0 ± 0.4	99.3 ± 0.6	99.1 ± 0.6
Alignment on ARS-UCD2.1				
Uniquely mapped reads (%)	33.9 ± 1.5	34.3 ± 1.6	35.8 ± 2	35.0 ± 3.1
Number of covered CpGs (million)	3.2 ± 0.1	3.2 ± 0.09	3.3 ± 0.1	3.3 ± 0.1
Average coverage per CpG	23.2 ± 3.6	24.0 ± 3.5	24.1 ± 3	22.9 ± 1.8
Percentage of CpGs10	60.2 ± 3.9	60.9 ± 3.5	62.2 ± 2.9	60.3 ± 1.9
Average DNA methylation at CpGs10 (%)	46.2 ± 1.6	46.5 ± 1.2	47.5 ± 1.7	47.8 ± 1.8
Percentage of hypomethylated CpGs10	50.9 ± 2	50.4 ± 1.2	49.3 ± 2	49.0 ± 2.1
Percentage of intermediate CpGs10	6.0 ± 1.1	6.0 ± 0.4	5.8 ± 0.6	6.1 ± 0.2
Percentage of hypermethylated CpGs10	43.4 ± 2.4	43.7 ± 1.1	44.9 ± 1.7	44.9 ± 1.9
Alignment on RepBase				
Uniquely mapped reads (%)	20.6 ± 1.7	20.4 ± 1.2	21.6 ± 1.9	22.4 ± 1.6
Number of covered CpGs (million)	1104 ± 72	1124 ± 85	1080 ± 38	1048 ± 4
Average coverage per CpG	36,996 ± 6451	37,722 ± 6606	43,479 ± 8507	43,603 ± 6136
Percentage of CpGs10	70.8 ± 3	70.8 ± 3.1	73.5 ± 2	73.3 ± 2
Average DNA methylation at CpGs10 (%)	30.8 ± 3	31.1 ± 3.2	31.3 ± 2.1	30.6 ± 2.9
Percentage of hypomethylated CpGs10	50.4 ± 2.7	49.5 ± 2.9	49.7 ± 2.9	48.4 ± 5.4
Percentage of intermediate CpGs10	41.3 ± 2	42.1 ± 1.8	40.8 ± 2.5	43.1 ± 5.4
Percentage of hypermethylated CpGs10	8.2 ± 1.9	8.5 ± 2.4	9.4 ± 1.7*	8.5 ± 1.5*

Values are mean ± standard error of the means. CpGs10: CpGs covered by at least 10 uniquely mapped reads. Hypermethylated, intermediate and hypomethylated CpGs10 indicate CpGs10 with average methylation percentages > 80%, [20%; 80%], and < 20%, respectively. Fertile and subfertile bulls were compared in both cohorts, and the asterisks indicate a significant difference in the independent cohort regarding the percentage of hypermethylated CpGs10 (Wilcoxon test, $p < 0.05$)

bias during RRBS library preparation or sequencing that might have affected subsequent results.

The average methylation at CpGs10 was 46.5%, which is consistent with values previously obtained on bovine sperm using RRBS [43] and also with the overall undermethylation reported for bovine sperm when compared to adult somatic cells using whole genome bisulfite sequencing [47]. Accordingly, a large proportion of the CpGs10 were hypomethylated (50.4% of CpGs10 with DNA methylation below 20% for the reference genome and 49.9% for the Repbase genome). Except for the percentage of hypermethylated CpGs10 in the Repbase genome, which was slightly reduced in the four subfertile samples from the independent cohort, there were no statistically significant differences between fertile and subfertile bulls in terms of the percentage of CpGs10 that were hypomethylated (DNA methylation below 20%), intermediate (DNA methylation between 20 and 80%) and hypermethylated (DNA methylation higher than 80%), thus indicating that subfertility is not related to global DNA methylation changes (Table 1).

In line with this finding, hierarchical clustering run on the DNA methylation values at CpGs10 failed to segregate samples according to bull fertility (Additional file 1: Fig. S1A). The main cohort included bulls from two semen collection centers and was split into several batches for semen processing and library preparation because of its huge size; we therefore also confirmed that neither the bulls' origin nor technical artefacts affected the DNA methylation patterns (Additional file 1: Fig. S1B–D).

Taken together, these results show that high quality RRBS data could be obtained regarding DNA methylation analysis in sperm, enabling the investigation of differences related to fertility in the largest cohort so far constituted in cattle.

Identification of fertility-related differentially methylated CpGs and regions

In order to determine any differences in DNA methylation between the fertility groups in the main cohort, 1,949,735 CpGs10 covered in at least 22 samples per group (which represents half of the smallest group) were subjected to differential analyses using two types of algorithms (see "Methods"). While no DMCs were identified using DSS, which is described as dealing with intra-group inter-individual variability [48], 3252 DMCs were obtained using methylKit [49]. Because sequence polymorphism is an important source of inter-individual variability in DNA methylation patterns [50], this striking difference in the number of DMCs obtained using either DSS or methylKit called for a careful examination of the genomic sequences imputed in the 100 bulls. The presence of polymorphisms affecting C and/or G in the CpG was indeed observed in 2352 out of 3252 methylKit DMCs, suggesting that in our dataset, this algorithm tended to select genetic rather than epigenetic variations. At these sites, the methylation differences between fertility groups were probably emphasized by a biased distribution of genotypes, resulting in a marked intra-group inter-individual variability that precluded their identification as DMCs by DSS but not by methylKit.

To suppress the polymorphisms affecting CpGs that could interfere with DMC detection, all putative variants recorded in the 1000 Bull Genome database and targeting the CpGs covered by RRBS were filtered out, resulting in a background of 1,548,563 CpGs10 covered in at least 22 samples per fertility group (22 samples representing half of the smallest group). Using methylKit, 490 DMCs and 46 differentially methylated regions (DMRs) were identified, which contrasts with no DMCs being obtained with DSS, and suggests that sources of intra-group inter-individual variations still existed in our dataset. This was also visible from the hierarchical clustering run on the CpGs10 before and after the filtering of sequence variants, which demonstrates that inter-individual epigenetic variations unrelated to fertility shape the sperm methylome independently of the presence of polymorphisms (Additional file 1: Fig. S1A, E). These inter-individual epigenetic variations were clearly not due to the presence of variable amounts of somatic cells in the samples (Additional file 1: Fig. S2).

Although a certain degree of overlap existed between the fertile and subfertile bulls, they

(See figure on next page.)

Fig. 2 Discrimination of fertile and subfertile bulls based on fertility-related differentially methylated CpGs and regions. **A**, **B** principal component analysis (**A**) and correlation clustering (**B**) run on the DMCs identified after variant filtering between fertile (red) and subfertile (blue) bulls in the main cohort. **A** Although a certain degree of overlap exists between the groups, they clearly segregate along dimension 1. Confidence ellipses are represented. **B** The cluster on the left mostly contains subfertile bulls, while the cluster on the right mostly contains fertile bulls. **C** Heatmap showing the average DNA methylation values at 18 DMRs covered in all 100 samples. Each cell of the heatmap is colored according to the average methylation value in the corresponding sample (displayed in columns, with fertile and subfertile bulls shown in red and blue, respectively) and DMR (in rows, with DMRs hyper- and hypomethylated in subfertile bulls shown in red and black, respectively). For each DMR, the genomic coordinates and the gene containing the DMR (if any) are indicated on the right-hand side. Three different clusters of DMRs were obtained manually. Cluster 1 included 5 DMRs that were highly methylated in subfertile bulls, while cluster 2 included 9 DMRs with low methylation in fertile bulls

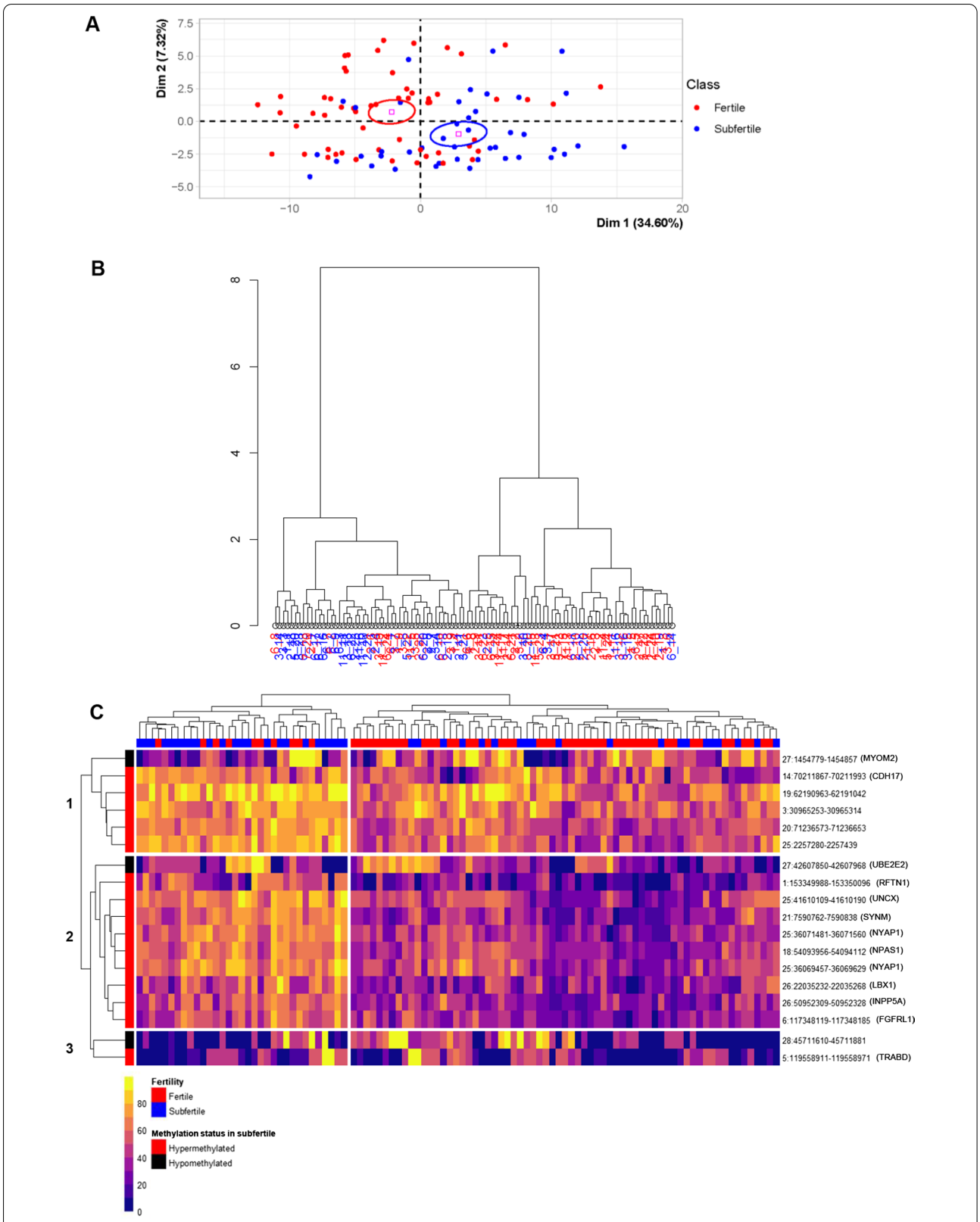


Fig. 2 (See legend on previous page.)

segregated as two distinct groups in a principal component analysis (PCA, Fig. 2A) and in hierarchical clustering run on the DMCs found after variant filtering (Fig. 2B). Likewise, among 18 DMRs covered in all 100 samples of the main cohort, the average DNA methylation status enabled the clustering of samples according to fertility on a heatmap (Fig. 2C). To verify that the overlap between fertile and subfertile bulls was not due to remaining polymorphisms directly targeting DMCs, the imputed genomic sequences in the 100 bulls were examined at DMCs. Only 21 over 490 DMCs (and 5 over the 107 DMCs without missing values used in the PCA) showed the presence of a polymorphism affecting the C and/or the G. Although we cannot definitely rule out that indirect genetic mechanisms could influence DNA methylation at DMCs and lead to this overlap, this result demonstrates that the 1000 Genomes filtering strategy efficiently suppressed most of the variations of methylation artificially due to variants affecting the DMCs. All together, these results suggest that despite the presence of marked inter-individual variability unrelated to fertility at the level of CpGs10, epigenetic information relevant to fertility was embedded in the 490 DMCs and 46 DMRs.

Fertility-related differentially methylated CpGs and regions target distinct genome features according to their methylation status in subfertile bulls

To characterize the 490 DMCs from a genomic point of view, their location was first examined as a function of their methylation status in subfertile bulls versus fertile bulls. Most DMCs were hypermethylated in subfertile bulls (386 out of 490; Additional file 4: Table S3), and these hypermethylated DMCs were found on every chromosome analyzed in the genome (Fig. 3). They were not distributed uniformly but scattered along regions with a dense accumulation of the background CpGs10 (represented by a grey line in Fig. 3), thus indicating no specific enrichment. By contrast, although the DMCs hypomethylated in subfertile bulls only accounted for 21% of all DMCs (104 out of 490), they clustered in discrete regions such as the extremities of chromosome 10 (Fig. 3), which could indicate the presence of restricted domains that tend to lose DNA methylation in subfertile bulls.

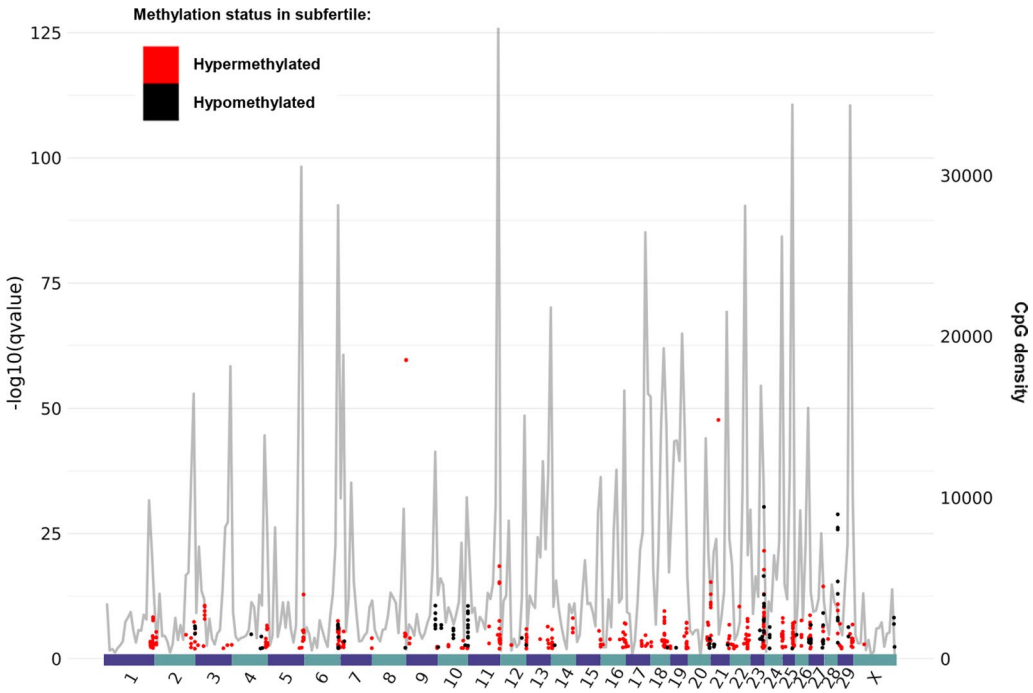


Fig. 3 Genomic location and methylation status of fertility-related differentially methylated CpGs. Black and red dots represent DMCs that are hypo- and hypermethylated in subfertile bulls, respectively. The left *y*-axis indicates a log function of the methylKit *q*-value that reflects the significance of the DMC. Chromosomes are displayed on the *x*-axis. The grey line shows the density of CpGs10 analyzed by RRBS (number of CpGs10 analyzed per window, each window corresponding to the length of the chromosome divided by 300; right *y*-axis). While DMCs hypermethylated in subfertile bulls are scattered throughout the regions targeted by RRBS, hypomethylated DMCs concentrate at discrete and specific regions of the genome

Consistent with the fact that most DMCs were hypermethylated in subfertile bulls, 38 out of 46 DMRs were also hypermethylated in subfertile bulls (Additional file 5: Table S4). At the 18 DMRs covered in all 100 samples of the main cohort, most subfertile bulls exhibited a DNA methylation value higher than 50% (orange to yellow, Fig. 2C) while the DNA methylation value of most fertile bulls was close to or below 50% (purple and blue). Of note, among the few DMRs that were hypomethylated in subfertile bulls, two were located on chromosome 10 where stretches of DMCs can be seen in Fig. 3 (Additional file 5: Table S4).

The DMCs and DMRs were annotated relative to gene features, repetitive elements and CpG islands, shores and shelves (Additional files 4 and 5: Tables S3 and S4), and the distribution of these different genome features in hypo- and hypermethylated DMCs was analyzed relative to the background. There were only small differences between the DMCs hypermethylated in subfertile bulls and the background (Fig. 4A, upper and middle panels), such as an enrichment in intergenic sequences, in long interspersed nuclear elements (LINEs) and in shelves; in parallel, there was a depletion of promoters, transcription start sites (TSSs), 5′ untranslated regions (5′UTRs) and CpG islands (CGIs). These results were quite similar to the genome features targeted by the 46 DMRs, most of which were made up of DMCs hypermethylated in subfertile bulls (Additional file 1: Fig. S3). The differences between DMCs hypomethylated in subfertile bulls and the background were more marked (Fig. 4A, upper and lower panels). Indeed, there was a clear enrichment of DMCs in intergenic sequences and depletion in exons, promoters and TSSs, which paralleled an overrepresentation of DMCs present in open sea relative to regions dense in CpGs (CGIs, shores and shelves). A striking enrichment in repetitive elements, and particularly in tandem repeats and LINEs, was also observed among these hypomethylated DMCs.

To further analyze these repeats, a differential analysis was performed between fertile and subfertile bulls using the methylation values obtained after alignment on the Repbase genome (Additional file 6: Table S5). This approach failed to highlight any CpGs10 displaying significant differential methylation with respect to bull fertility. However, individual CpGs10 located in the consensus LINE L1 consistently exhibited lower DNA methylation in subfertile bulls when compared to fertile bulls (Fig. 4B). Furthermore, the average DNA methylation of LINE L1, but not of LINE BovB, tended to be slightly decreased in subfertile bulls (Additional file 1: Fig. S4A), thus supporting the finding that DMCs hypomethylated in subfertile bulls are enriched in LINEs. The principal reason why no CpG in LINE L1 was found to be differentially methylated was due to the methylation difference between groups (2.7% on average), which was lower than the threshold of 10% set for the methylKit analysis. The same approach was applied to other families of repetitive elements referenced in Repbase, but no differences could be found between fertile and subfertile bulls, as exemplified by members of the long terminal repeat (LTR) family (Additional file 1: Fig. S4B).

To conclude this section, DMCs and DMRs hypermethylated in subfertile bulls accounted for most of the methylation differences found between the two fertility groups, and targeted genome features distinct from those targeted by hypomethylated DMCs and DMRs. Interestingly, LINEs were enriched in both hypermethylated (33% vs. 16% LINEs in the background) and hypomethylated (38% vs. 16%) DMCs. However, at least for LINEs in the L1 family, a trend toward hypomethylation in subfertile bulls was observed using a Repbase genome that has the potential to capture information from the multimapped reads.

Differentially methylated genes are related to development and sperm physiology

The next step was to analyze the biological function of the genes targeted by differential methylation between fertile and subfertile bulls. For this purpose, a functional enrichment analysis and an extensive review of the literature were conducted on the 170 genes containing at least one DMC in genic regions and/or in upstream

(See figure on next page.)

Fig. 4 Fertility-related DMCs target different genome features according to their methylation status in subfertile bulls. **A** The background CpGs10 and DMCs were annotated as described in the Methods. The percentage of each genome element in three functional genome features (genes, repetitive elements and CpG islands) is indicated on the different pie charts. DMCs were split into hypo- and hypermethylated DMCs according to their methylation status in subfertile bulls compared to fertile bulls. TSS: transcription start site, TTS: transcription termination site, UTR: untranslated region, upstream region: up to − 10 kb from the TSS, downstream region: up to + 10 kb from the TTS. **B** DNA methylation percentages (y-axis) were obtained for 60 individual CpGs10 included in the consensus sequence of the LINE L1 element (*x*-axis) after the alignment of RRBS sequences on a Repbase artificial genome. Each dot represents one sample, with fertile and subfertile bulls shown in red and blue, respectively. Red and blue lines indicate the trend over the 60 CpGs in fertile and subfertile bulls, respectively, and were obtained using the geom_smooth function of the ggplot2 R library. The 95% confidence interval is shown in light grey. The L1 element was slightly but consistently less methylated in subfertile bulls than in fertile bulls at all CpG positions

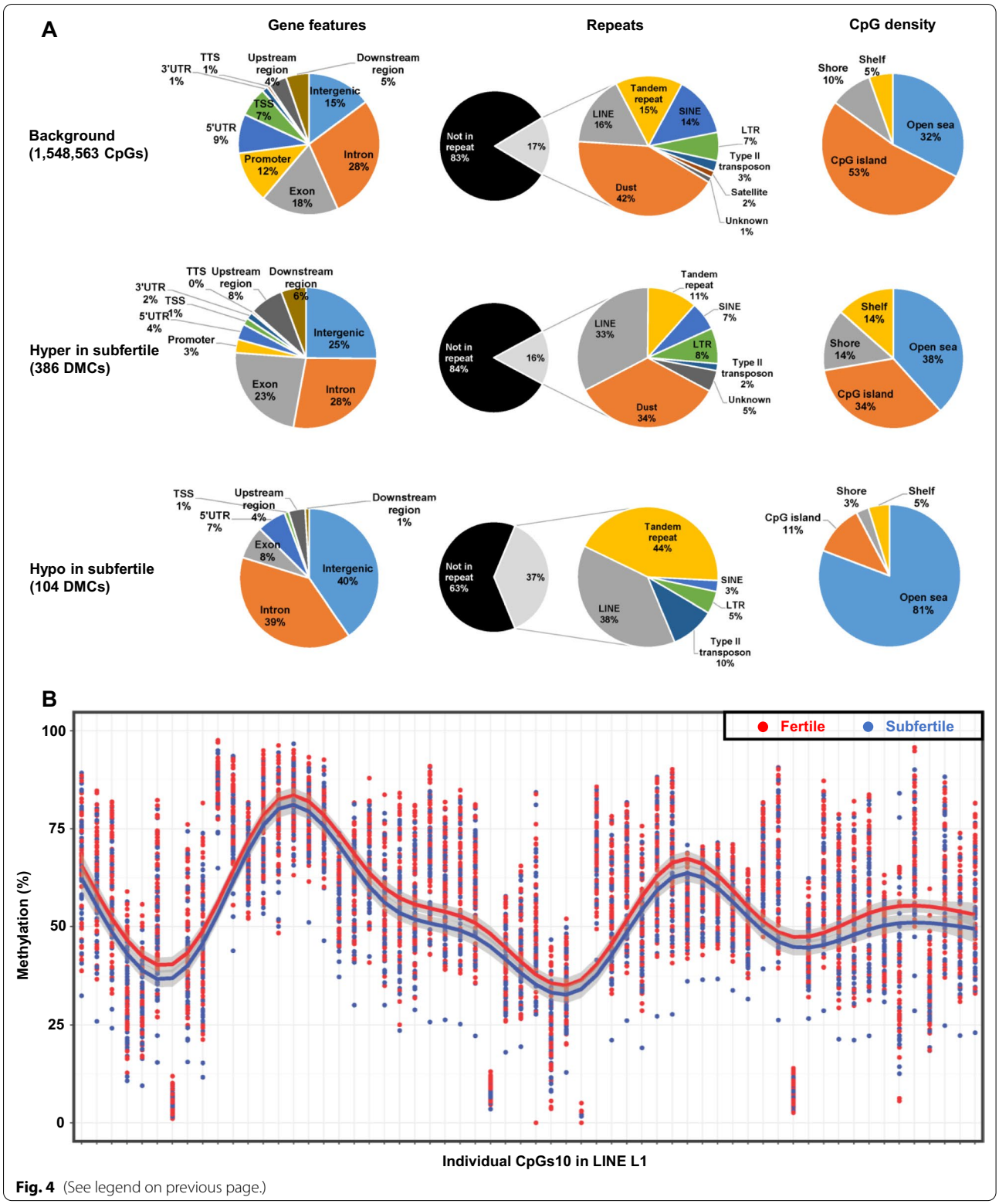


Fig. 4 (See legend on previous page.)

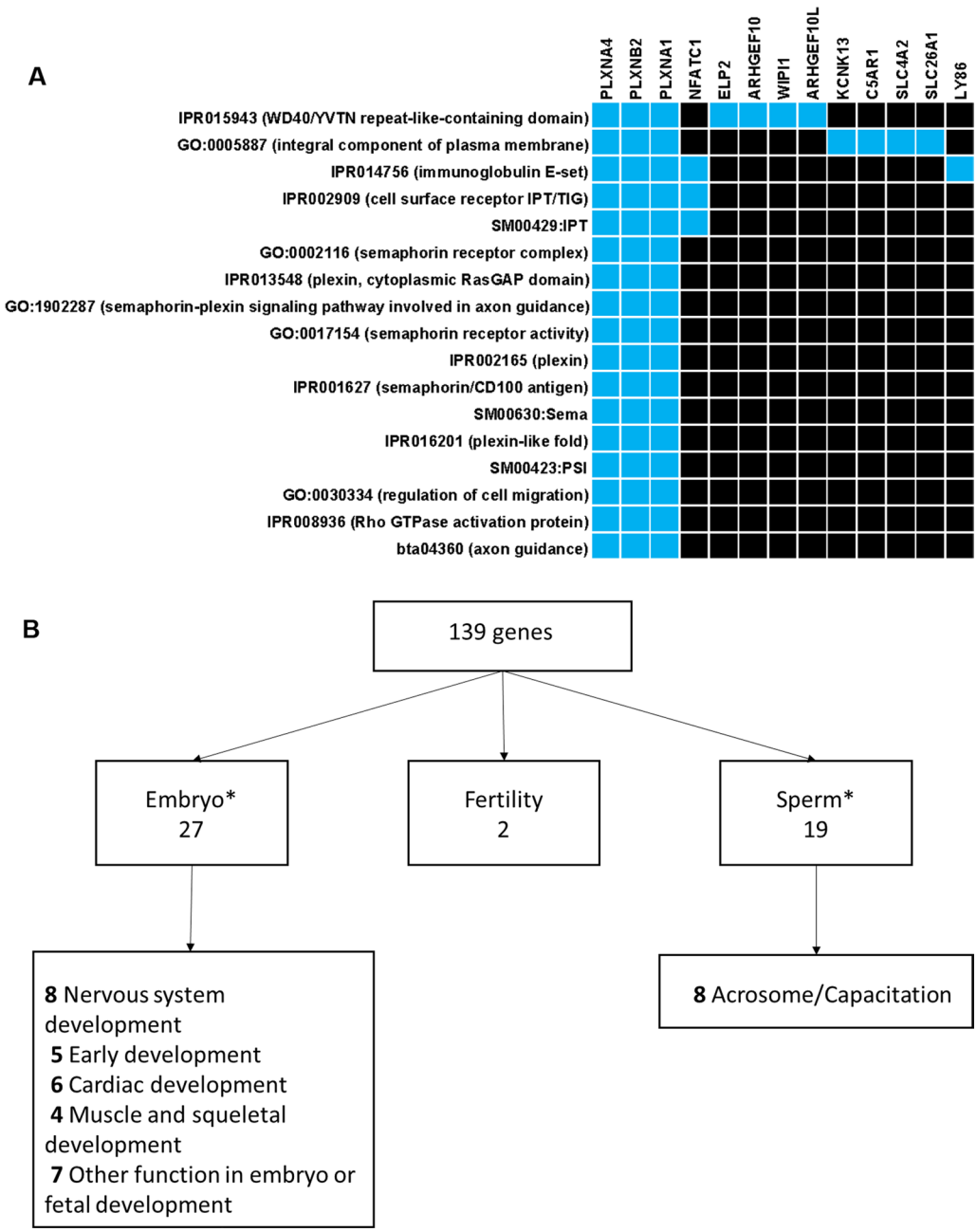


Fig. 5 Differentially methylated genes are relevant to male fertility. **A** functional enrichment analysis was performed using DAVID, where genes containing DMCs (170 genes) were compared to those covered in the background CpGs10 (19,829 genes). The enrichment cluster with an EASE score of 1.84 is represented, with genes in the cluster displayed on the x-axis and the different terms on the *y*-axis. A correspondence between a gene and a term is indicated by a light blue color on the heatmap. **B** A systematic review of the literature was undertaken for all genes associated with DMCs that contained a gene name ($n = 139$). The NCBI Pubmed database was interrogated using the gene names successively associated with each of the following terms: "embryo*", "sperm*" and "fertility". Some genes matched with several terms and are represented in several categories, explaining why the total number of counts exceeded the number of unique genes of interest (46 unique genes)

or downstream regions (up to 10 kb from a gene). To increase the number of genes in the analysis at this step, all DMCs were considered independently of their hypo- or hyper-methylation status in subfertile bulls. A Database for Annotation, Visualization and Integrated Discovery (DAVID) analysis highlighted a unique cluster of enrichment that reached significance (EASE score of 1.84, 1.3 being the lower limit for significance; Fig. 5A).

Three genes located on three different chromosomes, *PLXNA4* (chromosome 4), *PLXNB2* (chromosome 5) and *PLXNA1* (chromosome 22), which belong to the Plexin family, were associated with all the terms in the cluster. Plexins are present on the surface of cells and interact with the semaphorin family of proteins to trigger a rearrangement of the cytoskeleton and to regulate cell shape, differentiation, junctions, motility and survival. All these processes are involved in the remodeling of epithelia and endothelia during organogenesis as well as in axon guidance during development of the nervous system [51].

To go further, we performed a systematic review of the literature targeting the differentially methylated genes. Genes relevant to development, sperm function and fertility accounted for a large share of all differentially methylated genes (46 unique genes out of 139; 33%) and are listed, together with the corresponding references, in Additional file 7: Table S6. The largest share of these 46 unique genes was involved in embryonic and fetal development (Fig. 5B). It is noteworthy that genes involved in nervous system development were highly represented, which was consistent with the results of the DAVID analysis. By contrast, only five genes were related to early embryonic development, which concern cytokinesis, blastocyst formation and gastrulation. As for sperm function, genes related to the formation of the acrosome and capacitation represented eight out of 19 genes. Finally, two genes were associated with male and female fertility in genome-wide association studies in cattle.

We next compared the 139 genes differentially methylated between fertile and subfertile bulls with genes differentially methylated between cases and controls in four human studies investigating male fertility [26–28, 52]. Twenty-five differentially methylated genes were found in common with at least one human study, among which 18 genes exhibited a consistent methylation status between our study and the human studies (Additional file 8: Table S7). Out of these 25 genes, 17 were not highlighted during the systematic literature mining we performed (Additional file 7: Table S6). Because they were differentially methylated as a function of fertility in two different species, these 17 genes may be additional candidates relevant to male fertility.

In conclusion, differential methylation between fertile and subfertile bulls targeted genes with reported functions in sperm physiology, differentiation and post-testicular maturation, early and post-implantation development and processes important to organogenesis and nervous system development, all being relevant to male fertility.

Validation by bisulfite pyrosequencing

The quantification of DNA methylation using RRBS is reliant on the counts of reads bearing "C" and "T", which can lead to a certain degree of imprecision and to sampling biases, especially for CpGs where coverage is limited [53]. The quantification of DNA methylation at some CpGs using an independent technique was therefore necessary to validate both the molecular and bioinformatics aspects of the RRBS results.

Here, 11 DMCs located in five genes relevant to development (*PLXNB2, NPAS1, LBX1* and *SORCS2*) and sperm function (*ATG7*) were validated using bisulfite pyrosequencing (Fig. 6). These DMCs are located in diverse genomic contexts, such as gene upstream regions (*LBX1,* Additional file 1: Fig. S5), exons (*NPAS1*), introns (*SORCS2*), intron–exon junctions (*PLXNB2*) and 3′ UTRs (*ATG7*), and reflected the overall hypermethylation of DMCs in subfertile bulls when compared to fertile bulls (Fig. 6A). Because the main cohort was too large to be fully analyzed with the low-throughput system used for pyrosequencing, ten samples with contrasting levels of methylation were selected in each fertility category for each gene (Additional file 9: Table S8). For all genes, the average DNA methylation values at DMCs obtained using RRBS and bisulfite-pyrosequencing were significantly correlated (Spearman's correlation coefficient

(See figure on next page.)

Fig. 6 Bisulfite-pyrosequencing validations on twenty fertile and subfertile semen samples. **A** IGV browser views of the regions targeted for pyrosequencing in the *ATG7, NPAS1, SORCS2, LBX1* and *PLXNB2* genes. The red and blue bar charts represent the methylation percentages at each CpG10 position for fertile ($n = 10$) and subfertile ($n = 10$) bulls, respectively. The CpGs analyzed by pyrosequencing are numbered according to their 5′–3′ position along the genome. The CpGs identified as fertility-related DMCs are indicated in black text and red boxes, while non-DMCs are indicated in grey. **B** For each gene region, the average methylation percentage measured by pyrosequencing (*y*-axis) was calculated for the DMCs included in the region and plotted against the average methylation percentage measured by RRBS at the same DMCs (*x*-axis). Each dot represents one sample from the fertile (in red, n = 9 to 10) and subfertile (in blue, $n = 9$–10) groups. The least squares lines of best fit and Spearman's rank R correlation coefficients are indicated. All correlations were highly significant (Spearman's rank correlation test; $p < 0.05$). **C** Methylation percentages of individual CpGs assayed by pyrosequencing in fertile (in red, $n = 10$) and subfertile (in blue, $n = 10$) bulls. CpGs are numbered according to **A** and DMCs are highlighted in black. Dots show the methylation levels of individual samples, while the trends per fertility group are indicated by red and blue lines obtained using the geom_smooth function of the ggplot2 R library, together with 95% confidence intervals in light grey. Asterisks indicate that the methylation percentage measured by pyrosequencing differed significantly between fertility groups for all analyzed CpGs (Wilcoxon test, $p < 0.05$)

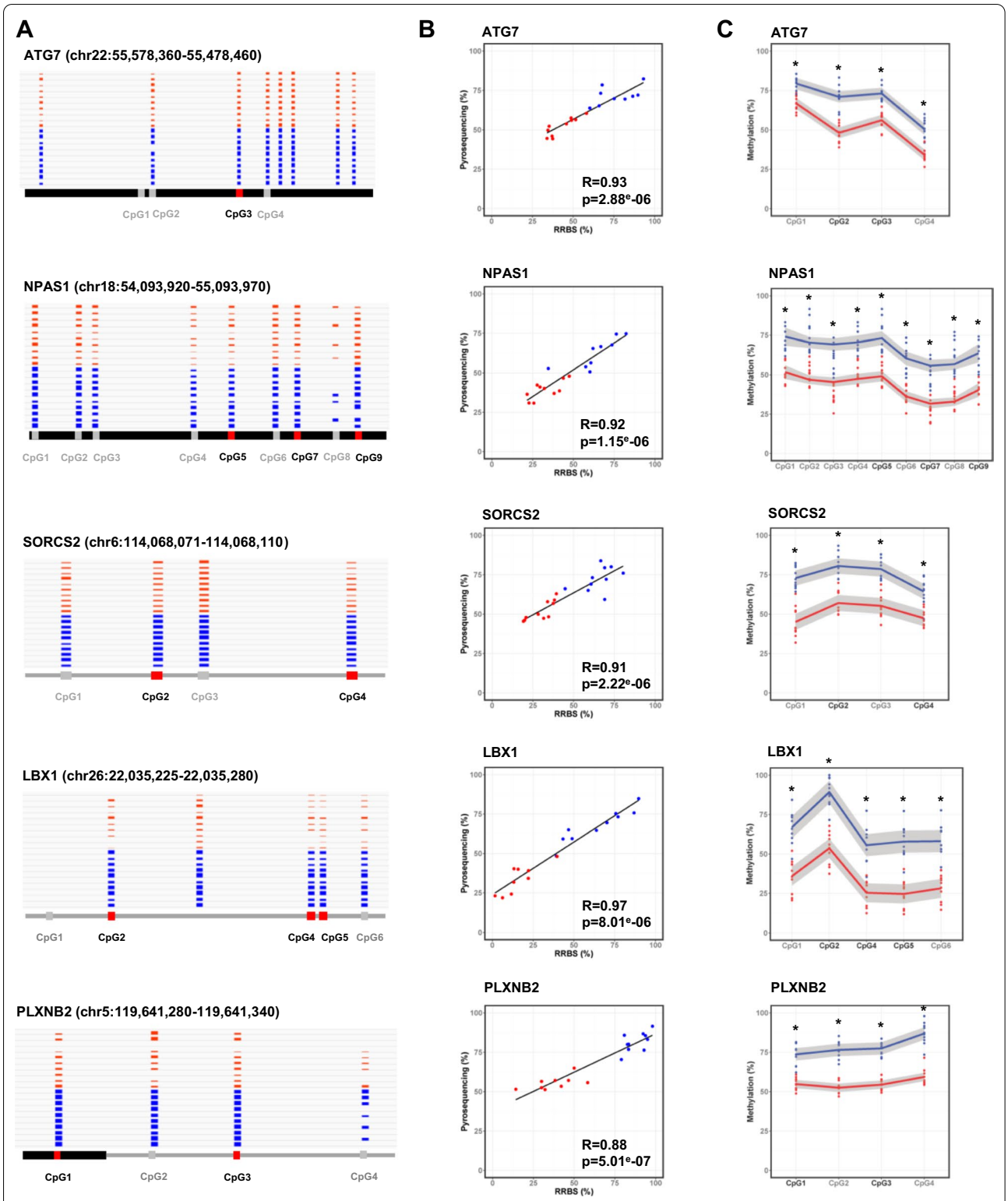


Fig. 6 (See legend on previous page.)

between 0.88 and 0.97; $p < 0.05$), thus validating our RRBS data from a technical perspective (Fig. 6B). As expected, pyrosequencing confirmed the significantly higher DNA methylation level in subfertile bulls at all the DMCs investigated (Fig. 6C).

Pyrosequencing enabled the quantification of 15 additional CpGs surrounding DMCs. Interestingly these CpGs displayed the same behavior as nearby DMCs, being significantly more methylated in subfertile bulls in the pyrosequencing data (Fig. 6C). This trend could also be seen in the RRBS data (Fig. 6A) and was in line with the many reports showing that CpGs located within the same region tend to behave in the same way [54–56]. Despite the similar behavior of nearby CpGs, of the five genes we analyzed more closely only *LBX1* and *NPAS1* contained a DMR; more generally, only 225 out of 490 fertility-related DMCs clustered into DMRs. These results may be linked to the maximal inter-DMC distance required to constitute a DMR (see "Methods"), which could be a limitation under an RRBS approach that covers discontinuous portions of the genome [57]. Another reason may have been the insufficient coverage of nearby CpGs which were therefore not integrated into the background and did not undergo differential analysis.

Taken together, the bisulfite-pyrosequencing results (1) were strongly aligned with the RRBS findings; (2) confirmed the hypermethylated status of DMCs in subfertile bulls compared to fertile bulls for five genes relevant to male fertility, and (3) suggest that the number of DMRs identified during our analysis may have been underestimated because of the stringency of the bioinformatics settings. Fertility-related DMCs could therefore be used with confidence in subsequent steps of our study.

Bull fertility status can be predicted from the sperm methylome using a Random Forest approach

Bull fertility was next modelled using the DNA methylation values obtained on the main cohort at 107 DMCs with no missing values (Additional file 4: Table S3). For the model construction and validation, the cohort was split randomly into a training set (2/3 of the animals) and a testing set (1/3) with the same proportion of fertile and subfertile bulls as in the original dataset (Fig. 7A). Using a Random Forest approach, which is described as suitable for molecular biology data [58], a model was then built using the training set, and the prediction was assessed by comparing the estimated fertility and actual fertility of the testing set. The average performances of the model were calculated from 50 iterations of this process, with resampling of the training and testing sets at each iteration. This cross-validation indicated satisfactory performance, with values for the area under the receiver operating characteristics (ROC) curve (AUC) and accuracy of 0.80 and 0.72, respectively (Fig. 8A). The model displayed higher sensitivity (0.80) than specificity (0.63), which means that errors were most frequently caused by a misclassification of subfertile bulls.

Because the sperm methylome is sensitive to a wide range of environmental variations [59] there is a potential risk that DNA methylation at fertility-related DMCs may vary beyond fertility, which could alter the predictive performance of the model that was developed using these DMCs. The main cohort included bulls that were commercialized by two breeding companies, maintained in two different semen collection centers 100 km distant from each other and under different animal management practices. This experimental design offered an opportunity to assess the degree to which the performance of the model was affected by the origins of the bulls (Fig. 7B). Using the 107 DMCs with no missing values, a new Random Forest predictive model was built from the methylation values measured on the 56 samples collected in center 1 and tested on the 44 samples collected in center 2 (Fig. 8B), and vice-versa (Fig. 8C). Satisfactory performance was achieved in both cases, with accuracies of 0.77 and 0.83, and AUCs of 0.84 and 0.76, respectively. Sensitivity and specificity varied more markedly and in opposite directions. Indeed, the model trained on the samples collected in center 1 and tested on samples collected in center 2 was best for the correct prediction of subfertile bulls (specificity of 0.75). By contrast, the model trained on the samples collected in center 2 and tested on samples collected in center 1 failed to correctly predict subfertile bulls (specificity of 0.55) but outperformed in terms of the correct prediction of fertile bulls (sensitivity of 0.92). This result thus demonstrated that the model's performance was only moderately impacted by the origins of the bulls, suggesting the robustness of the model regarding a certain degree of environmental variation.

In order to understand why 20–30% of the bulls were systematically misclassified in our model, PCA was run on the DNA methylation values measured at DMCs, and individuals were classified as fertile bulls predicted as being fertile (FF), fertile bulls predicted as being subfertile (FS), subfertile bulls predicted as being subfertile (SS) and subfertile bulls predicted as being fertile (SF). Very similar results were observed with the three models: fertile and subfertile bulls that were correctly predicted were discriminated by the first dimension of PCA, while misclassified bulls tended to cluster with the opposite fertility class (Fig. 8D–F). This result clearly showed that at DMCs, the 20–30% misclassified bulls displayed a DNA methylation signature typical of the opposite class, thus producing the wrong predictions. Several factors were therefore investigated in order to relate these

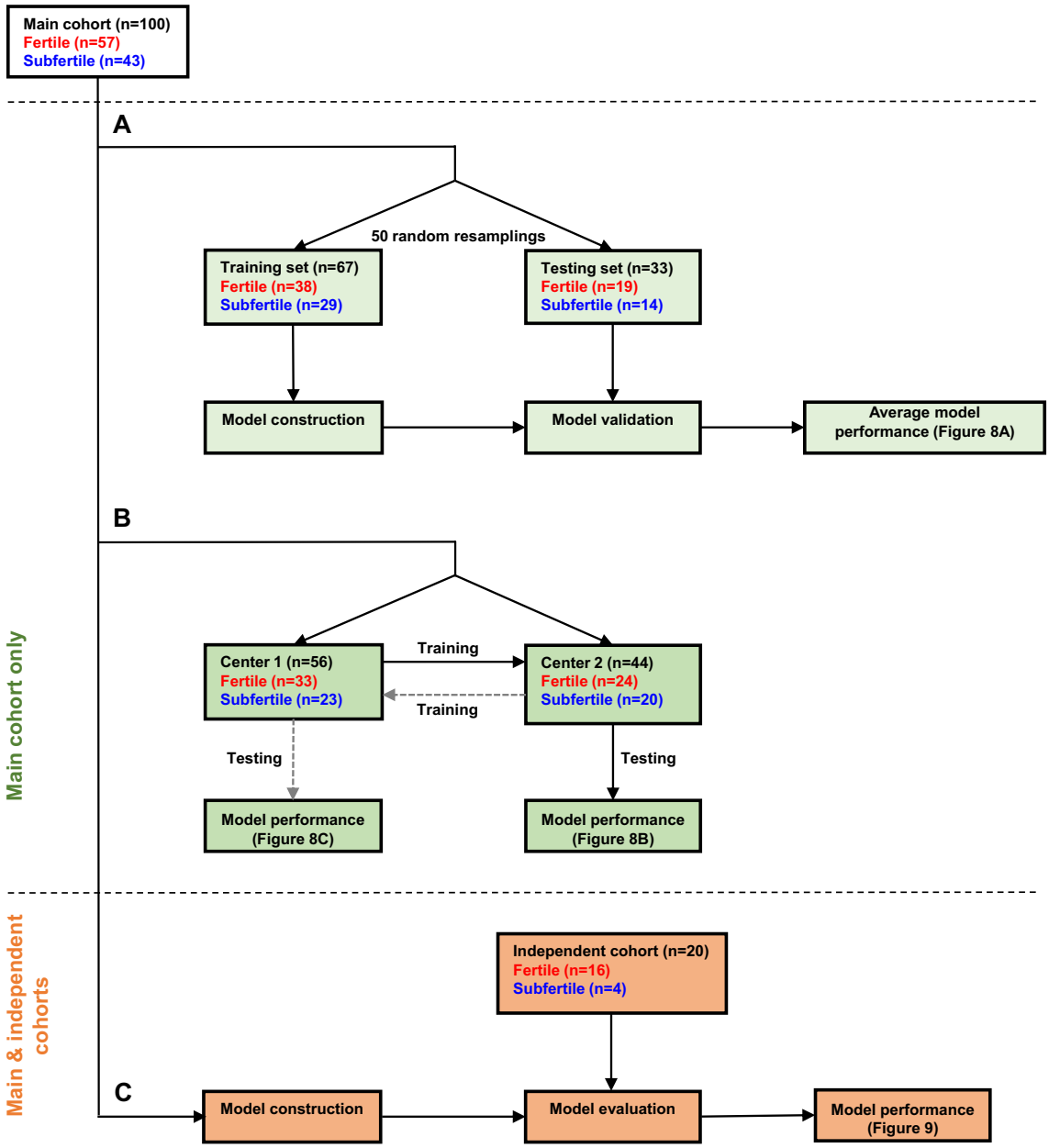


Fig. 7 Strategies to evaluate the performance of models predictive of fertility status. **A** model construction and cross-validation. The main cohort was split into a training set and testing set. The training set contained 2/3 of the samples ($n = 67$) randomly selected but with conservation of the original proportion of fertile and subfertile samples, and was used to create the model. The testing set contained 1/3 of the remaining samples ($n = 33$) and was used to evaluate the predictive ability of the model previously constructed on the training set. This process was iterated 50 times with a resampling of the training and testing sets, and the model performance was averaged over these 50 iterations. **B** Effect of the bulls' origin on the model performance. The main cohort was split into a semen collection center 1 set ($n = 56$) and a semen collection center 2 set ($n = 44$), which were then used alternatively to train and test the predictive model. **C** Potential of the model for field application, using an independent and less controlled cohort. The whole main cohort was used to train the model, which was then tested on the independent cohort

unexpected DNA methylation patterns to biological or technical features specifically affecting the misclassified bulls. No significant differences between misclassified and correctly classified bulls in both fertility classes were observed regarding fertility (Additional file 1: Fig. S6A), which indicated that although they displayed a DNA methylation pattern that mimicked that of fertile bulls, the misclassified subfertile bulls were indeed subfertile;

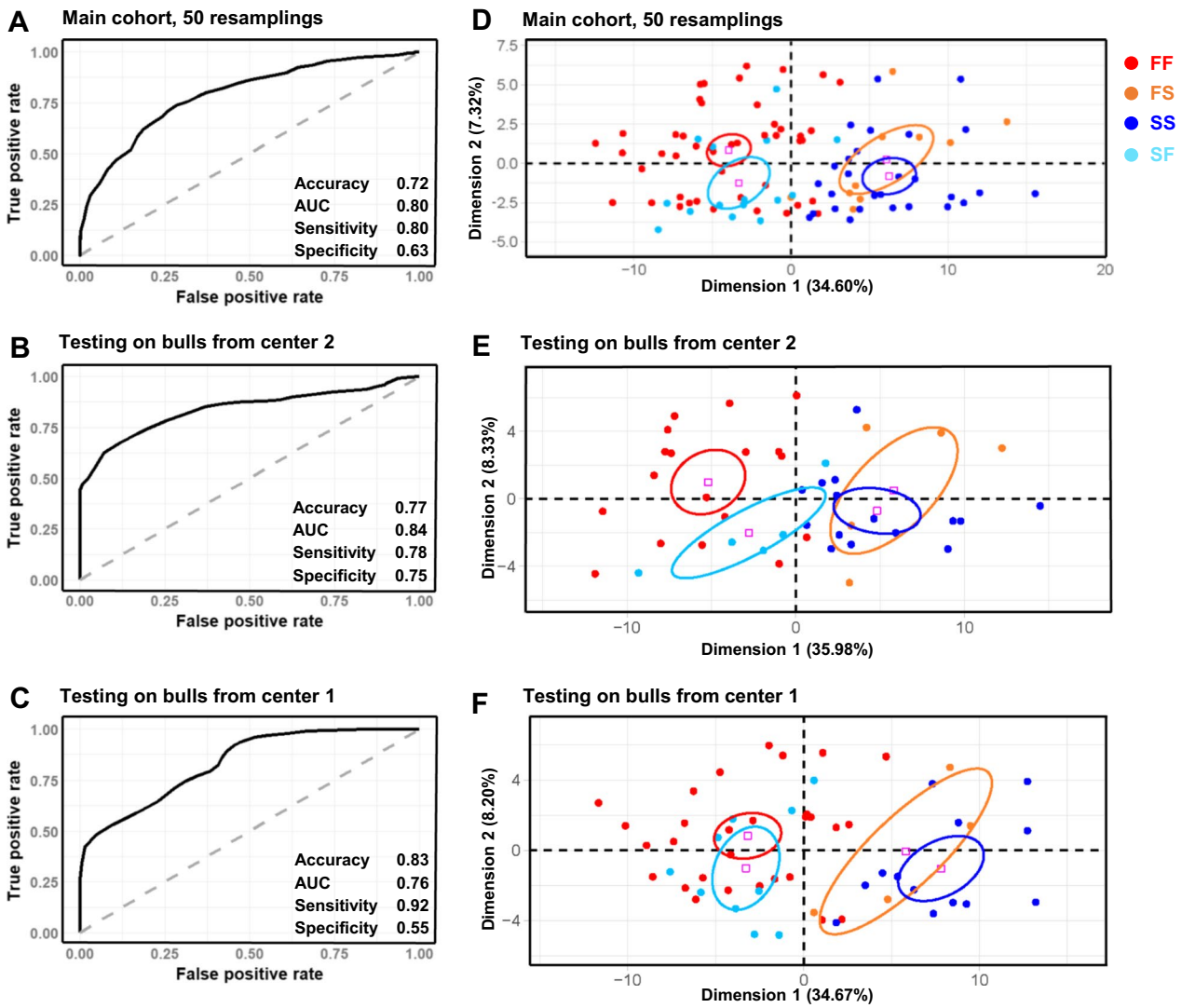


Fig. 8 Fertility in the main cohort can be predicted from DMCs. **A**–**C** Random Forest models were built on three different training sets from the DNA methylation values of 107 fertility-related DMCs without missing values. The models were then run on three different testing sets, and predicted fertility status was compared against the actual fertility status of bulls included in each type of testing set. Receiver operating characteristics (ROC) curves and four performance indicators are displayed: accuracy (correct prediction rate), area under the ROC curve (AUC), sensitivity (true positive rate) and specificity (true negative rate). **D**–**F** For each model, principal component analysis was run on the 107 fertility-related DMCs and an individual factor map is displayed, with bulls colored according to the consistency between their predicted and actual fertility status (red: fertile predicted as fertile, FF; blue: subfertile predicted as subfertile, SS; orange: fertile predicted as subfertile, FS; light blue: subfertile predicted as fertile, SF). Confidence ellipses are indicated. **A**, **D** the main cohort was split into a training set of 67 bulls and a testing set of 33 bulls, the proportions of fertile and subfertile bulls being the same in each set. The average performance indicators were calculated after 50 iterations with a random resampling of the training and testing sets. **B**, **E** The training set included bulls originating from semen collection center 1 while the testing set included bulls originating from semen collection center 2. **C**, **F** The training set included bulls from center 2 while the testing set included bulls from center 1. Whatever the model, the performance indicators were within the same range; misclassified bulls were found in both fertility classes and exhibited a DNA methylation pattern typical of the opposite class

the reverse was also found for misclassified fertile bulls. Moreover, a later fertility indicator (sire conception rate, SCR) further confirmed their correct assignment to the two fertility classes (Additional file 1: Fig. S6B), which was initially performed according to the NRR 56 (Additional file 1: Fig. S6A). No significant differences between the FF, FS, SF and SS groups were found regarding the number of AIs used to measure fertility (Additional file 1: Fig. S6C) and semen functional parameters (Additional file 1: Fig. S6D–F). The four groups were also

equivalently distributed in terms of the season of semen collection, number of ejaculates per sample and technical batches (Additional file 1: Fig. S6G–J), thus demonstrating the absence of confounding factors in our experimental design. In the absence of any identified source of bias, it is therefore possible that the fertility of misclassified bulls was independent from DNA methylation at these 107 DMCs. To extend the panel of CpGs, we tried to impute missing values and built a model using 295 DMCs, but this did not cause any significant change to the results (Additional file 1: Supplementary Results and Additional file 1: Fig. S7).

The performance of the model was next assessed on bulls in the independent cohort that had not been used to identify DMCs (Fig. 7C). It should be noted that this independent cohort was less controlled in terms of the age of the bulls, which ranged from 15 to 39 months, and each bull was only represented by one ejaculate, since the pooling of several ejaculates would clearly limit the practical applications of our study. The accuracy and AUC of the model on this independent cohort (respectively 0.72 and 0.78 in Fig. 9A) were comparable to those found using the main cohort, suggesting that practical applications of the model could be envisioned in field conditions. Unlike the results described above, sensitivity was, however, lower than specificity, which indicates that the misclassifications resulted from fertile bulls predicted as being subfertile. A PCA was run on the methylation values of the independent cohort and, as previously observed with the main cohort, FS bulls clustered separately from FF bulls (Fig. 9B), suggesting again that the fertility of misclassified bulls was independent of DNA methylation. Unlike the main cohort, no misclassification was observed for subfertile bulls. This result may have been related to the unbalanced number of bulls in each class ($n = 4$ subfertile bulls vs. $n = 16$ fertile bulls), which was closer to the incidence of subfertility in the whole population than that seen in the main cohort.

Taken together, these results demonstrate that for approximately 75% of the bulls, field fertility could consistently be predicted from the sperm DNA methylation status of 107 CpGs, whatever the origin and age of the bulls and the number of ejaculates in the sample. For 25% of the bulls belonging to both fertility classes, prediction was hampered by a reversed DNA methylation status at these CpGs, but also at an extended panel of 295 CpGs. In the absence of any identified source of bias, these results suggest that the fertility of a subset of bulls was independent of DNA methylation, at least at the CpGs investigated.

Discussion

This study was performed on the largest cohort so far used to investigate the relationships between male fertility and DNA methylation in cattle, and one of the

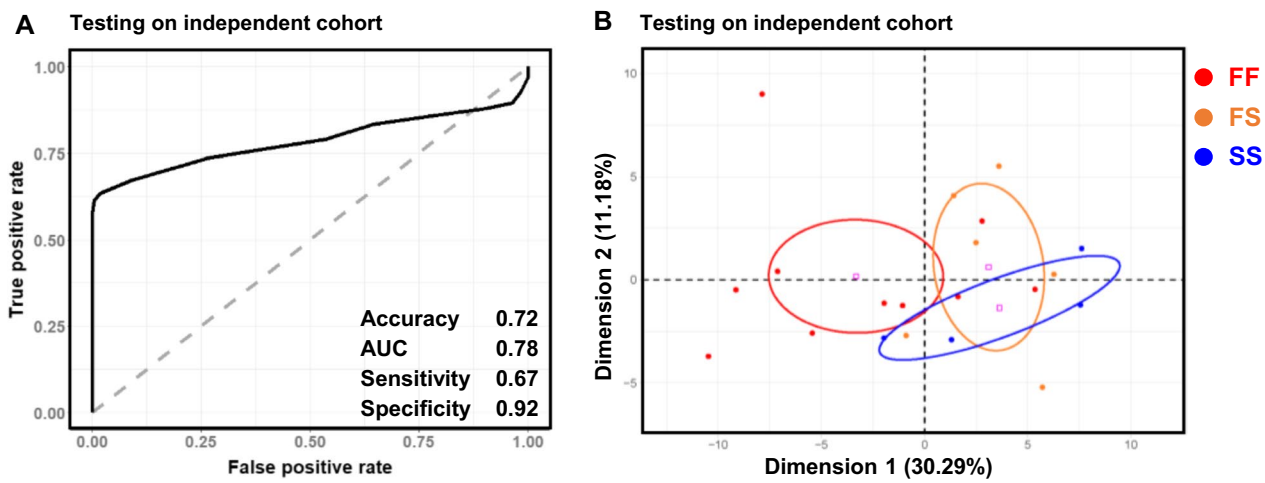


Fig. 9 Fertility in an independent cohort can be predicted from DMCs. **A** a Random Forest model was built on the main cohort using the DNA methylation values of 107 fertility-related DMCs without missing values. The model was then run on the independent cohort, and the predicted fertility status was compared with the actual fertility status. ROC curves, accuracy, AUC, sensitivity and specificity were comparable to the performance indicators obtained when the testing set was included in the main cohort. **B** principal component analysis was run on the 107 fertility-related DMCs, and an individual factor map is displayed, with bulls colored according to the consistency between their predicted and actual fertility status (red: fertile predicted as fertile, FF; blue: subfertile predicted as subfertile, SS; orange: fertile predicted as subfertile, FS). Confidence ellipses are indicated. Misclassified bulls were found in the fertile class only and displayed a DNA methylation pattern similar to that of the subfertile class

largest cohorts of any mammals, including humans. Field fertility was very precisely assessed for 120 bulls using two different indicators and we also demonstrated that the experimental design was devoid of any obvious confounding effects. Using this valuable resource, we generated high quality and technically validated genome-wide DNA methylation data using RRBS. Our main finding was that although they did not differ in terms of semen functional parameters at 17–19 months of age, fertile and subfertile bulls displayed subtle DNA methylation differences in the same semen samples, and these differences could be used successfully to build models predictive of fertility status for the whole career of the bulls.

Inter-individual variability of the sperm DNA methylome

One striking finding revealed by our analysis was the important inter-individual variability observed in the sperm methylome at both the genome-wide and DMC levels. This inter-individual variability, which could be appreciated for the first time in cattle thanks to the large size of the main cohort, was independent of fertility and impacted all the results of our study. It firstly precluded the identification of DMCs using a stringent algorithm such as DSS [48], leading to a certain degree of overlap between fertile and subfertile bulls at the DMCs identified using the less stringent algorithm methylKit [49]. Heterogeneity at DMCs then slightly altered the performance of the predictive model, as it led to the misclassification of 20–30% bulls displaying a DNA methylation profile typical of the opposite class.

Inter-individual variability was not related to variations affecting individual ejaculates, since each bull was represented by several ejaculates. Because the automation of DNA methylation assays has been reported to improve reproducibility [60], we used a partially automated process to generate libraries, and checked that technical artefacts did not confound our results at each step of data generation. The sequencing parameters thus varied within the same range for the two fertility groups, and the RRBS results were not affected by semen processing and library preparation batches. Moreover, the excellent correlations between the RRBS and bisulfite-pyrosequencing results demonstrated that the technical limitations inherent to RRBS did not account for inter-individual variability in our data, and confirmed that our study was adequately powered in terms of sample size and sequencing depth to detect small differences between groups [53]. Genetic factors are another important source of variation in DNA methylation patterns [50], and it has been proposed that both genetic factors and somatic cell contamination confound DNA methylation analyses in human sperm [31]. Although we cannot definitively rule out the possibility that genetics interfered with our results via indirect mechanisms, we limited this effect by filtering out putative variants from the CpGs we analyzed. We also checked that the DNA methylation results were not altered because of residual contamination by somatic cells.

Because the effects of most technical factors could be regarded as insignificant in our study, one possibility is that stochastic, indirect genetic effects or uncontrolled environmental or physiological factors underlie inter-individual variability in the sperm methylome [61]. Among the physiological factors that have been described to impact the sperm methylome, age has a major effect in cattle [41, 42], humans [62] and mice [63], but could be excluded here as only ejaculates collected from bulls aged between 17 and 19 months were used. Because the sperm epigenome is also responsive to a wide range of environmental factors [59], we confirmed that the bulls' origins with respect to semen collection center did not interfere with the RRBS results and model performance. However, bulls are usually maintained in various herds before being recruited by a breeding company, and these diverse conditions may have modified the sperm methylome as a function of each individual bull's environment, finally affecting fertility without inducing a homogenous DNA methylation signature. Male germ cells are indeed subject to intense epigenetic remodeling during in utero and early post-natal life, and the vulnerability of the methylome to environmental variations during these periods may alter fertility [11, 12, 64]. Another common practice that might also affect the sperm methylome is that once arrived in a station, male calves are fed in order to optimize their average daily gain. The sperm epigenome is sensitive to nutrition [9] and modifications to early life nutrition in cattle have been reported to induce persistent changes to the sperm methylome [65]. The metabolic response of each bull to this practice may therefore vary as a function of its genetics and initial body condition, leading to epigenetic inter-individual variability. Finally, because it has also been suggested that heat stress might alter bull fertility through epigenetic mechanisms [66], we investigated the effects of the season of ejaculate collection on our experimental design. The distribution of fertile and subfertile bulls in both hot and cold seasons, whatever their methylation pattern and the outcome of the prediction, means that a risk of heat stress confounding the overall results of our study is unlikely, but does not completely exclude that it contributed to a certain degree of inter-individual variability. Therefore, although we controlled numerous factors likely to interfere with fertility-related variations in the sperm methylome, uncontrolled factors still remained in our study. Because humans live under far less standardized conditions and

display more genetic diversity than cattle, it is probable that the magnitude of inter-individual variability is even more important in the human sperm methylome. Of all the confounding factors present in humans, the effect of this epigenetic inter-individual variability on the consistency of the results reported on male fertility may have so far been overlooked [29, 32].

Prognostic value of the sperm DNA methylome for male fertility

Despite the important inter-individual variability we observed in our methylation data, the majority of the bulls displayed a DNA methylation profile at DMCs that enabled the creation of a predictive model with satisfactory performance. Importantly, this model displayed comparable performance regardless of the bulls' origin, and also when tested on individual ejaculates from an independent cohort that included bulls of various ages, thus demonstrating its robustness to certain variations in environmental or physiological factors. Interestingly, fertility status could be successfully predicted for all individuals with a DNA methylation pattern typical of their fertility class at DMCs, which demonstrated the potential of Random Forest approaches to model phenotypic outcomes from DNA methylation.

To our knowledge, this study represents the most comprehensive attempt to model bull fertility from the sperm methylome, which was probably enabled by the large size of the main cohort. Inadequate sample size has indeed been proposed as a major obstacle to replicating fertility-related DNA methylation signatures in human sperm [29]. In line with this view, the important inter-individual variability we report here makes it unlikely that the methylation differences previously reported between small numbers of bulls of contrasting fertility ($n = 3$–9 per group; [35–40]) can reflect those in a larger population with less extreme differences in fertility, thus limiting modelling approaches adapted from these datasets. Only Takeda et al. [39] confirmed the differential methylation between low and high fertility bulls using a wider population ($n = 50$) at 10 loci and were able to construct a predictive model that has now to be assessed on an independent cohort.

In our model, 20–30% of bulls with divergent DNA methylation profiles were consistently misclassified whatever the testing set used, underscoring the importance of a thorough assessment of models using independent populations. Of note, comparable model performance was also achieved using an extended panel of DMCs after the imputation of missing values. From a statistical point of view, this observation suggests that whatever the number of variables in the model, DNA methylation at DMCs does not suffice to explain the whole variance related to fertility. It is possible that the relatively limited differences in fertility that exist between fertile and subfertile bulls, together with the important inter-individual variability we report, may have precluded the identification of more discriminant DMCs from which the percentage of misclassifications would have been weaker. Importantly, both fertile and subfertile bulls were found to be misclassified, which indicates that a DNA methylation profile typical of fertile bulls is not sufficient to be fertile, which might be expected given the multifactorial nature of fertility. More surprisingly however, a DNA methylation profile typical of subfertile bulls does not necessarily lead to subfertility, suggesting the existence of compensatory mechanisms. In line with this finding, a study investigating the association between the sperm methylome and fecundity in humans reported that only 54% of men displaying the unfavorable DNA methylation pattern actually failed to conceive [28]. To improve predictive performance, a future direction might be to build the model using a less biased selection of CpGs than that resulting from a between-group differential analysis, as DNA methylation patterns at fertility-related DMCs are not conserved in all individuals. A further step would be to integrate other types of epigenetic features that might capture complementary aspects of the variance related to male fertility.

Biological features targeted by differential methylation

From a biological perspective, the comparable model performances obtained on the main and independent cohorts demonstrated that at DMCs, fertility-related variations of DNA methylation could be replicated in a significant proportion of animals that were completely independent from DMC identification. The biological features associated with these DMCs therefore represent a valuable source of information that could help to improve our understanding of male fertility.

Among the 139 genes targeted by DMCs, 19 were found to be important for sperm physiology, differentiation and post-testicular maturation, which is obviously of considerable interest regarding fertility. For instance, *ATG7*, on which we focused during the bisulfite-sequencing validation, is involved in autophagy and associated with formation of the acrosome and with spermiogenesis [67]; impairment of these functions in *ATG7*−/− germ cell-specific mice drive complete infertility [68]. Although several genes related to acrosome function were found to be differentially methylated in our study, it is unlikely that the subfertile bulls suffered from major spermiogenesis defects, which would probably have led to a more severe phenotype than that observed. However, in light of the phenotype described in *ATG7*−/− germ cell-specific

mice, it would be interesting to analyze the acrosome function in these bulls in greater detail.

During our study, 27 genes involved in development, 8 of them related to nervous system development, were also found to be differentially methylated between fertile and subfertile bulls. This finding agreed well with the absence of major alterations to semen functional parameters in subfertile bulls, as impaired development could lead to loss of the embryo and subsequently to a reduction in fertility without necessarily affecting semen functional parameters. We focused in particular on three genes, which were further analyzed by bisulfite-pyrosequencing: *NPAS1*, *LBX1* and *SORCS2*. *NPAS1* is a member of the basic helix-loop-helix per-ARNT-SIM (bHLH-PAS) family of transcription factors. Several genes in this family are known to be involved in nervous system development [69]; *NPAS1* in particular has been related to fertility in a genome-wide association study conducted on cattle [70]. *SORCS2* encodes a VPS10 domain-containing receptor that is highly expressed in developing neural tissues [71, 72], and *LBX1* is a homeobox gene orthologous to the *ladybird* gene in Drosophila, which is important for limb, neural and heart development [73–76]. A DAVID analysis also highlighted three genes involved in axon guidance*: PLXNA1, PLXNA4, PLXNB2* (which we further analyzed by bisulfite-pyrosequencing). Strikingly, these three genes encode plexins of different classes that are located on different genome regions, suggesting the functional importance of this protein family to bull fertility. Plexins are receptors of semaphorins and both types of protein belong to the semaphorin signaling pathway that regulates cell adhesion, migration, division, differentiation and survival, acting on a wide range of developmental processes [51]. *PLXNB2* in particular is highly expressed in neuronal progenitors and its knockout leads to severe brain malformations and to developmental arrests before birth [77, 78]. Interestingly, the relative abundance of genes involved in neuron differentiation and axon guidance has also been reported by two studies which compared the sperm methylome of fertile and infertile men [26, 79]. Because selection of the bulls in our study was based on the NRR 56, differences in developmental outcomes up to 56 days of gestation may contribute to subfertility. Gestational stages until 56 days are critical for neural tube patterning in cattle [80]; methylation changes affecting genes involved in this process may therefore compromise the viability of the embryo or fetus, hence affecting NRR 56.

Twenty-five differentially methylated genes were found in common between our study and four human studies [26–28, 52], and the same biological functions related to sperm physiology and development were affected in previous studies published on bull fertility [35–40]; thus strengthening the relevance of these genes and functions to male fertility. Another biological feature that appeared to be conserved in other studies conducted on fertility-related variations in the sperm methylome was that most differentially methylated loci were hypermethylated in subfertile or infertile cases when compared to controls in bulls [40], boars [81] and humans [28, 31, 52, 79, 82]. Because sperm cells are hypomethylated relative to somatic cells in many species, including humans [83] and cattle [43, 47], it has been proposed that the genome-wide erasure of DNA methylation was impaired during the differentiation of male germ cells in infertile men with spermatogenesis defects [82]. More recently, hypermethylation at DMRs has been attributed to a larger proportion of contaminating somatic cells in sperm samples from oligozoospermic patients [31]. However, these two hypotheses are unlikely to apply to the present study, since we demonstrated that hypermethylation at DMCs and DMRs in subfertile bulls was not due to residual somatic cells, and that subfertile bulls did not suffer from obvious spermatogenesis defects. The functional role of DNA methylation in gene expression cannot be predicted as a whole, because it varies as a function of the gene elements targeted by differential methylation [84]. Furthermore, many distant regulatory elements such as enhancers are still poorly characterized in cattle and will fall into the intergenic class, which is enriched among DMCs independently of their methylation status in subfertile bulls. The absence of significant transcriptional activity in sperm cells is a further complication encountered during functional analyses of the sperm methylome. Considering all these limitations, it can be speculated that the DNA methylation differences we detected in transcriptionally silent sperm cells may reflect suboptimal transcriptional activity, either earlier during a bull's life for genes involved in sperm differentiation and physiology, or later after fertilization with respect to developmental genes, both processes resulting in subfertility.

In contrast with the overall hypermethylation of DMCs/DMRs in subfertile bulls, we observed a strong enrichment in LINE retrotransposons among hypomethylated DMCs, as well as a trend toward hypomethylation for LINEs in the L1 family but not the BovB family. Interestingly, the bovine genome contains more potentially active copies of L1 than of BovB retrotransposons; and consistent with this, L1 repeats also seem to have arisen more recently during evolution than the BovB family [85]. De novo DNA methylation guided to L1 repeats by PIWI-interacting RNAs (piRNAs) offers a safeguard against the mobilization of retrotransposons in the germline, and disruption of this defense mechanism in male mice leads to genome invasion and meiotic arrest [86],

suggesting that the demethylation of L1 repeats might be at risk regarding the genome integrity of cattle germ cells. The hypomethylation of L1 has been reported in the testes of patients with spermatogenic failure, and associated with the down-regulation of genes in the piRNA pathway, suggesting that the molecular mechanisms involved in the epigenetic silencing of retrotransposons were not fully functional in these patients [87]. In line with this hypothesis, L1 expression has been claimed to increase in the testes of patients with impaired spermatogenesis [88]. However, other studies have not reported any significant changes to L1 DNA methylation in the sperm cells of infertile men [27, 30], which is also consistent with the small degree of difference we observed between fertile and infertile bulls. The expression of L1 repeats needs to be tightly regulated not only in germ cells but also after fertilization, where it regulates global chromatin accessibility in mouse embryos [89]; a loss of L1 silencing mechanisms in germ cells or embryos has also been proposed to cause early spontaneous miscarriage in humans [90]. Given the absence of severe spermatogenic defects in subfertile bulls, it is tempting to speculate that the slight hypomethylation we observed in sperm may lead to a suboptimal level or timing of the expression of L1 repeats after fertilization rather than to massive genome invasion during germ cell differentiation. This suboptimal expression may alter the dynamics of post-fertilization reprogramming, ultimately resulting in developmental arrest or pregnancy losses.

Conclusion

Using the largest cattle cohort in the field and extensive DNA profiling analyses, we were able to demonstrate that the bull sperm methylome displays important inter-individual variability. This inter-individual variability remains poorly investigated in humans and deserves further attention, as it may reflect earlier exposures that could be passed to the next generation. We identified a facultative DNA methylation signature that predisposed to subfertility and from which fertility status for the whole career of the bulls could be predicted consistently in at least 70% of bulls from different breeding companies and of different ages. Our results are promising in terms of applications, but also suggest the existence of independent and/or compensatory mechanisms for the regulation of male fertility. Identifying these mechanisms will offer further opportunities regarding the practical use of these biomarkers to predict fertility.

Methods

Bull cohorts and semen samples

Semen samples from 120 French Montbéliarde bulls, grouped in two independent cohorts, were used for this study. All samples were prepared from frozen semen straws commercialized for AI and stored in liquid nitrogen.

The main cohort was used to generate DNA methylation data using RRBS, identify fertility-related DMCs and construct and validate the predictive model. This included 100 bulls born between 2011 and 2014 and commercialized by the breeding companies Umotest and Evajura, and maintained in two different semen collection centers located 100 km distant from each other in France ($n = 56$, center 1 and $n = 44$, center 2), 57 of which were classified as fertile and 43 as subfertile. Fertility was defined at the bull level based on the non-return rate at 56 days post-insemination (NRR 56), obtained as follows. For each bull, the AI outcomes were obtained from all the AIs performed using the semen of this bull in 2017, 2018 and 2021. Each AI was given a score of 0 if another AI of the mated cow was observed within the 56-day subsequent interval, and 1 otherwise. To eliminate any bias due to the spurious association with other factors, the bull fertility indicator was estimated with the linear model applied to the 0/1 score of all AIs in the population, and used in the French bovine genetic evaluation of female fertility for selection purpose [91]. This model included the fixed effects of herd-year, month-year, parity of the cow, interval between calving and insemination, week day, AI technician, category of semen (sexed vs. conventional), and the random effects of genetic and permanent environmental effects of the cow, and of the bull. The bull effect was assumed to be normally distributed with zero mean and variance equal to 0.01 phenotypic variance. The bull effect estimate was used in the present study and referred to as "corrected NRR 56". To obtain one semen sample, 8–10 straws per bull that represented 2 to 5 ejaculates collected at 17–19 months of age and within a short period of time (6–52 days), were pooled after thawing for 30 s at 37 °C. Subsamples of 15 µL were used immediately to analyze semen functional parameters. The remaining semen was centrifuged for 7 min at 3500 g at room temperature and the extender was removed. Although microscopic examination did not reveal any significant contamination by somatic cells, the cell pellets were washed once with H_2O. Unlike somatic cells, bull spermatozoa are resistant to this treatment, which therefore guaranteed the absence of any residual somatic cells in the samples. After a further wash in phosphate buffer saline (PBS 1×), the sperm pellets were resuspended in PBS 1× and divided into aliquots for various experimental purposes. The equivalent of one straw (20 million sperm cells) was used for DNA extraction.

The independent cohort was used to produce RRBS data and evaluate the potential of the model built on the main cohort for field applications. This included 20 bulls

born between 2009 and 2012 and marketed by the breeding company Umotest. One ejaculate was analyzed per bull, whose age at semen collection ranged from 15 to 39 months. Ejaculates were defined as fertile ($n=16$) or subfertile ($n=4$) based on the corrected NRR 56 calculated for each bull and as explained above for the main cohort. One straw was used to assess semen functional parameters and extract DNA, as described above.

The entire process, from straw thawing to semen assessment and DNA extraction, was conducted in 7 batches of 2–24 samples per batch for the main cohort, and 7 batches of 1–6 samples per batch for the independent cohort. For each sample, information on the batch, the bull (year of birth, semen collection center), ejaculates (dates of collection, number of straws in the sample), the corrected NRR 56 and SCR are provided in Additional file 2: Table S1. Corrected SCR was obtained as described above for corrected NRR 56, except that each AI was given a score of 0 if no birth was reported for the mated cow after the expected gestation length, and 1 otherwise, as described by Barbat et al. [91].

Semen functional parameters

Semen motility was assessed by computer-assisted semen analysis (CASA, IVOS II, Hamilton Thorne, IMV Technologies). Five µL out of 15 µL pooled semen were mixed with 10 µL Easy buffer B (IMV Technologies) and incubated for 5 min at 37 °C. After incubation, 4 µL of this mix was loaded into a standardized four-chamber counting slide (Leja), which was then placed into the CASA system. Sperm motility was averaged from 10 microscope fields analyzed using the predetermined starting point set within each chamber. Results were expressed as percentages of motile sperm in the sample. Sperm viability and mitochondrial status were assessed from 5000 cells using the easyCyte 8HT flow cytometer and CytoSoft software (Guava, IMV Technologies). For sperm viability, membrane integrity was assessed using EasyKit 1 Viability and Concentration (IMV Technologies) that stains viable spermatozoa with intact membranes in green. Results were expressed as percentages of viable sperm in the sample. Sperm mitochondrial status was assessed using the EasyKit 2 (IMV Technologies) and expressed as the percentage of polarized mitochondria that appeared in orange fluorescence. Both kits were used as described by Sellem et al. [92].

In order to account for variations between the different series of experiments, the data were corrected for the batch effect using a linear model, with the sample preparation batch (Additional file 2: Table S1) set as a fixed effect. Residuals of this model were extracted for further analyses. The results obtained for viability, mitochondrial status and motility, with and without correction for the batch effect, are provided in Additional file 3: Table S2 and Additional file 1: Fig. S8.

Genomic DNA preparation

Approximately 20 million spermatozoa prepared as described above were pelleted and resuspended in 200 µL lysis buffer (10 mM Tris–HCl pH 7.5, 25 mM EDTA, 1% SDS, 75 mM NaCl, 50 mM dithiothreitol and 0.5 µg glycogen), and incubated overnight at 55 °C in the presence of 0.2 mg/ml proteinase K. After incubation with 25 µg/ml RNAse A for 1 h at 37 °C, genomic DNA was extracted twice using phenol:chloroform (1:1) and chloroform, then precipitated with ethanol and washed. The dried pellet was resuspended in TE buffer (10 mM Tris HCl pH 7.5, 2 mM EDTA) and the DNA concentration was measured using a Qubit 2.0 Fluorometer with the dsDNA BR Assay kit (Invitrogen). The integrity of genomic DNA was confirmed for all samples by agarose gel electrophoresis.

Reduced representation bisulfite sequencing

RRBS libraries were produced in 12 batches for the main cohort and 4 batches for the independent cohort. All pipetting steps before final amplification were carried out using an NGS STARlet liquid handling system with four channels (Hamilton), ensuring reproducibility between the different library preparation batches. Genomic DNA (200 ng) was digested with MspI, end-repaired and ligated overnight with Illumina adapters [43, 93]. The following day, size selection was performed using SPRIselect magnetic beads (Beckman-Coulter) as previously reported [65]. The DNA was then converted twice with sodium bisulfite using the EpiTect bisulfite kit (Qiagen) following the manufacturer's instructions. Converted DNA was amplified with Pfu Turbo Cx hotstart DNA polymerase (Agilent) using 14 PCR cycles. The libraries were purified using AMPure XP beads (Beckman-Coulter) and DNA concentrations were measured with a Qubit 2.0 Fluorometer with the dsDNA HS Assay kit (Invitrogen). Electrophoresis on a 4–20% precast polyacrylamide TBE gel (Invitrogen) and staining with SYBR green confirmed the homogeneous pattern for all libraries, with fragments ranging from 150 to 400 bp (40–290 bp genomic DNA fragments + adapters). The libraries were finally sequenced on an Illumina HiSeq4000 sequencer to produce 75 bp paired-end reads (Integragen SA).

Bioinformatics analyses

On average, sequencing generated 33 and 35 million read pairs per library for the main cohort and independent cohort, respectively. The sequences displayed the expected nucleotide composition based on MspI digestion and bisulfite conversion according to FastQC quality

control (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Subsequent quality checks and trimming were carried out using TrimGalore v0.4.4 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), which removed adapter sequences, poor quality bases (Phred score below 20) and reads shorter than 20 nucleotides. The bisulfite conversion rate was estimated from unmethylated cytosine added in vitro during the end-repair step. This reached 99% on average (Table 1) and the minimal value for all 120 samples was 98.1%. High quality reads were aligned on the bovine reference genome (ARS-UCD2.1 assembly) using Bismark v0.20.0 in the default mode with Bowtie 1.2.1 [94, 95]. Only CpGs covered by at least 10 uniquely mapped reads (CpGs10) were retained for subsequent analysis. To avoid the confounding effects of sequence polymorphisms and bisulfite conversion, 401,172 CpGs10 that co-localized with a putative sequence polymorphism affecting the C and/or G were filtered out [96]. Since no sequence polymorphism information was available for unplaced scaffolds, CpGs located on unplaced scaffolds were also filtered out. Each remaining CpG10 was assigned a methylation percentage per sample calculated from Bismark methylation calling $\left(\text{number of reads with } \text{“}C\text{”} \times 100\right)/(\text{number of reads with } C + \text{number of reads with } \text{“}T\text{”})$, which could be visualized using the Integrative Genome Viewer (IGV) genome browser [97]. PCA and hierarchical clustering were then computed on the matrix of methylation percentages for each CpG10 with no missing values, using the FactoMineR R package [98]. For hierarchical clustering, the distance between samples was calculated using Pearson correlation coefficients and Ward's method was applied as the linkage function.

A subset of 1,548,563 CpGs10, covered in at least 22 samples per fertility group of the main cohort, constituted the background. The DMCs were identified from this background using methylKit v1.0.0 [49] and DSS v2.14.0 [48] in the default mode. With methylKit, a CpG10 was considered as a DMC when the adjusted p value (q value, obtained after SLIM correction) was lower than 0.01 and the average methylation difference between the two groups was at least 10% according to the methylKit calculation mode, which takes account of the coverage per sample. With DSS, the same minimal methylation difference applied, but the threshold for the adjusted p value was set at 0.1, and p value adjustment was performed according to the Independent Hypothesis Weighting method, using the alpha parameter set at 5% and the average methylation per group as a covariable [99]. DMRs were defined as regions containing at least 3 DMCs with an inter-distance between each DMC of 100 bp or less. It should be noted that both the DMC and DMR datasets contained missing values because of samples in which the coverage thresholds were not reached.

For Fig. 2C, an average DNA methylation value per DMR and per sample was calculated from all the CpGs10 included in the DMR. The heatmap was then generated for 18 DMRs displaying DNA methylation values for all 100 samples using the R package pheatmap (v1.0.12).

Annotation of the background CpGs10, DMCs and DMRs was performed as described [65] relative to gene features, CpG density and repetitive elements using an in-house pipeline. The reference files were downloaded from Ensembl (ftp://ftp.ensembl.org/pub; release 95). The following criteria were applied: TSS, − 100 to + 100 bp relative to the transcription start site (TSS); promoter, − 2000 to − 100 bp relative to the TSS; TTS, − 100 to + 100 bp relative to the transcription termination site (TTS); shore, up to 2000 bp from a CGI, and shelf, up to 2000 bp from a shore. A site/region was considered to belong to a CGI (respective shore and shelf) if an overlap of at least 75% was observed between the site/fragment and the CGI (respective shore and shelf). A site/region was considered as being overlapped by a repetitive element whatever the extent of this overlapping. The lists of fertility-related DMCs and DMRs with annotation features are provided in Additional files 4 and 5: Tables S3 and S4, respectively.

Genes containing DMCs in intragenic regions and/or in the upstream (up to − 10 kb from the TSS) and downstream (up to + 10 kb from the TTS) regions were subjected to an enrichment analysis using DAVID with default parameters [100]. The 19,829 genes covered by RRBS (i.e., all genes containing at least one background CpG10) were used as the reference. Clusters with an enrichment score above 1.3 were taken into account.

To better characterize repetitive elements, an artificial genome containing the consensus sequence of each bovine repeat was constituted from the Repbase database [46]. Reads were aligned on this artificial genome as explained above, and the average methylation rate was calculated per CpG10 for each repeat and each sample. A differential methylation analysis between fertile and subfertile bulls was then performed using methylKit [49], as described above.

Bisulfite pyrosequencing

For five genomic regions, ten samples per fertility group were selected based on their contrasting methylation patterns under RRBS analysis (Additional file 9: Table S8). Bisulfite conversion was performed on 0.5 µg genomic DNA using the EpiTect bisulfite kit (Qiagen). Primers were designed using the Qiagen Pyromark assay design software (Additional file 10: Table S9) and amplifications were performed using the Pyromark PCR kit (Qiagen)

according to the manufacturer's instructions. The following PCR program was used: 15 min at 95 °C followed by 45 cycles of 30 s at 94 °C, 30 s at 56 °C, 30 s at 72 °C, and finally 10 min at 72 °C. The reverse primers were 5′-biotinylated for all five regions. After denaturation and washes, the purified biotinylated strand of PCR products was employed as a template for pyrosequencing with 0.3 μM pyrosequencing primer, using the Pyromark Q24 device and Pyromark Gold Q96 reagents (Qiagen). Each CpG was analyzed in duplicate, and inconsistent duplicates (more than 5% difference) were repeated. For each sample, the DNA methylation percentage per CpG was obtained by calculating the mean of all consistent replicates that passed quality control by the Pyromark Q24 software. DNA methylation values obtained by pyrosequencing were compared between fertile and subfertile bulls using a Wilcoxon test. The significance of correlations between the average DNA methylation values obtained by RRBS and pyrosequencing for each region was tested using Spearman's rank correlation test.

Random Forest predictive model

The predictive model was built on the matrix of methylation percentages at DMCs, where samples were in rows and DMCs in columns. The DMCs with missing values were either filtered out before model construction or conserved if they contained less than 10% missing values; the latter were then imputed using the R package missMDA with default options [101]. Random Forest models were built using the 'rf' option from the R package caret (v6.0-84) [102]. The "mtry" parameter was estimated by the square root of the number of features in the model, and the number of trees was set at 500.

Three different strategies were applied to assess the performance of the predictive model (Fig. 7), and the predicted fertility status was compared with the actual fertility status of samples included in each type of testing set. For this purpose, ROC curves were computed [103], and four model quality indicators were calculated: model accuracy (correct prediction rate), AUC, sensitivity (true positive rate) and specificity (true negative rate). For the first strategy with 50 resamplings of the testing test, the average accuracy, AUC, sensitivity and specificity were considered to evaluate the model performance by cross-validation.

Other statistical analyses

The groups were compared using the non-parametric Wilcoxon test if the following criteria did not apply: (1) more than 15 values available per group; (2) normal distribution according to the Shapiro–Wilk test; (3) same variance of the two groups according to the F-test. Among a series of comparisons, the Wilcoxon test was used for all comparisons regarding consistency, even if some comparisons fulfilled the above criteria.

Literature mining

A systematic review of the literature was performed for all genes differentially methylated with the Biomart annotation available ("Gene name" column in Additional file 4: Table S3; 139 genes out of 170). The NCBI Pubmed database was interrogated using the gene names successively associated with each of the following terms: "embryo*", "sperm*" and "fertility", and all the relevant references were collected (Additional file 7: Table S6).

The comparison with human case studies was performed as follows: among all the listed references, those focusing on the comparison of genome-wide DNA methylation patterns between fertile and infertile/subfertile human sperm samples, and with accessible supplementary data available, were pointed out. This led to the selection of four studies highlighting 1843 [27], 31 [26], 3 [28] and 384 [52] differentially methylated genes. These four gene lists were compared with the 139 genes with the Biomart annotation available identified during the current study. The genes found in common are listed in Additional file 8: Table S7. For studies where the information was available, the methylation status in subfertile samples is indicated.

Abbreviations

AI: Artificial insemination; AUC: Area under the ROC curve; CASA: Computer-assisted semen analysis; CGI: CpG Island; DAVID: Database for annotation, visualization and integrated discovery; DMC: Differentially methylated CpG; DMR: Differentially methylated region; IGV: Integrative genome viewer; NRR 56: Non-return rates of inseminated cows at 56 days post AI; PBS: Phosphate buffer saline; PCA: Principal component analysis; ROC: Receiver operating characteristics; RRBS: Reduced representation bisulfite sequencing; SCR: Sire conception rate; TSS: Transcription start site; TTS: Transcription termination site; UTR: Untranslated region.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13148-022-01275-x.

Additional file 1: Fig. S1.. correlation clustering run on the methylation percentages at CpGs10 covered in at least 22 samples per group. **Fig. S2**: heatmap run on CpGs discriminant between sperm and somatic cells. **Fig. S3**: annotation of fertility-related DMRs relative to different genome features. **Fig. S4**: average DNA methylation at individual LINE and LTR repeats after alignment of the RRBS sequences on a Repbase artificial genome. **Fig. S5**: pyrosequencing validation of the upstream region of LBX1 gene. **Fig. S6**: fertility, semen functional parameters and semen sample characteristics in correctly classified and misclassified bulls. **Fig. S7**: PCA run on the methylation percentages at DMCs, with or without the imputation of missing values. **Fig. S8**: semen functional parameters before and after correction for the batch effect.

Additional file 2: Table S1. Listing the semen samples, information on their origin and processing, and the field fertility of the corresponding bulls.

Additional file 3: Table S2. Listing the functional parameters measured on each semen sample.

Additional file 4: Table S3. Listing the fertility-related DMCs, their DNA methylation status in each sample and their annotation regarding genome features.

Additional file 5: Table S4. Listing the fertility-related DMRs, their methylation status in each fertility group and their annotation regarding genome features.

Additional file 6: Table S5. Listing the CpGs10 obtained after alignment of the RRBS sequences on a Repbase artificial genome, their coverage and DNA methylation status in each sample.

Additional file 7: Table S6. Listing the differentially methylated genes relevant to fertility together with supporting references.

Additional file 8: Table S7. Listing the differentially methylated genes found in common with four human studies.

Additional file 9: Table S8. Listing the semen samples used for pyrosequencing validation.

Additional file 10: Table S9. Listing the primers used for pyrosequencing validation.

Acknowledgements
This research was made possible thanks to the semen samples kindly supplied by the breeding companies Umotest and Evajura. We are also grateful to the Genotoul bioinformatics platform Toulouse Midi-Pyrenees (Bioinfo Genotoul) for providing computing and storage resources, to Victoria Hawken for English editing and to Hélène Hayes for helpful suggestions.

Author contributions
VC analyzed data and drafted the manuscript. ACT and JPP generated the RRBS libraries. VC, ACT, ES, CP, CLD, LS, HJ and HK performed other experiments. AAF and LJ performed bioinformatics analyses. ES, CH and SF were involved in the experimental design. CH, MB, MPS and DB were involved in the genetic analyses. ES, CLD, LS, HJ and HK conceived the study. LS, HJ, FJ and HK obtained funding. FJ and HK supervised the study and edited the manuscript. All authors have read and approved the final version of the manuscript.

Funding
This study was funded by the French National Research Agency (Grant ANR-13-LAB3-0008-01 SeQuaMol) and APIS-GENE (AP-2018-44). VC was a CIFRE fellow of the French National Agency for Research and Technology (ANRT). The funding bodies had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript. APIS-GENE has read and approved the final version of the manuscript.

Availability of data and materials
Additional data files are provided (see above). RRBS fastq files have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB46371 (https://www.ebi.ac.uk/ena/data/view/PRJEB46371).

Declarations

Ethics approval and consent to participate
Not applicable (only commercial semen samples were used for the purposes of this study).

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]INRAE, BREED, Université Paris-Saclay, UVSQ, 78350 Jouy-en-Josas, France. [2]Ecole Nationale Vétérinaire d'Alfort, BREED, 94700 Maisons-Alfort, France. [3]R&D Department, ALLICE, 149 rue de Bercy, 75012 Paris, France. [4]Université Paris-Saclay, AgroParisTech, INRAE, GABI, 78350 Jouy-en-Josas, France.

Received: 23 December 2021 Accepted: 8 April 2022
Published online: 27 April 2022

References

1. Brugo-Olmedo S, Chillik C, Kopelman S. Definition and causes of infertility. Reprod Biomed Online. 2001;2:41–53.
2. Tournaye H, Krausz C, Oates RD. Novel concepts in the aetiology of male reproductive impairment. Lancet Diabetes Endocrinol. 2017;5:544–53.
3. Krausz C, Riera-Escamilla A. Genetics of male infertility. Nat Rev Urol. 2018;15:369–84.
4. Agarwal A, Baskaran S, Parekh N, Cho C-L, Henkel R, Vij S, et al. Male infertility. Lancet Lond Engl. 2021;397:319–33.
5. Carrell DT. Epigenetics of the male gamete. Fertil Steril. 2012;97:267–74.
6. Gannon JR, Emery BR, Jenkins TG, Carrell DT. The sperm epigenome: implications for the embryo. Adv Exp Med Biol. 2014;791:53–66.
7. Schagdarsurengin U, Paradowska A, Steger K. Analysing the sperm epigenome: roles in early embryogenesis and assisted reproduction. Nat Rev Urol. 2012;9:609–19.
8. Boissonnas CC, Jouannet P, Jammes H. Epigenetic disorders and male subfertility. Fertil Steril. 2013;99:624–31.
9. Schagdarsurengin U, Steger K. Epigenetics in male reproduction: effect of paternal diet on sperm quality and offspring health. Nat Rev Urol. 2016;13:584–95.
10. Ibrahim Y, Hotaling J. Sperm epigenetics and its impact on male fertility, pregnancy loss, and somatic health of future offsprings. Semin Reprod Med. 2018;36:233–9.
11. McSwiggin HM, O'Doherty AM. Epigenetic reprogramming during spermatogenesis and male factor infertility. Reprod Biosci Ltd. 2018;156:R9-21.
12. Marcho C, Oluwayiose OA, Pilsner JR. The preconception environment and sperm epigenetics. Andrology. 2020;8:924–42.
13. Cannarella R, Condorelli RA, Mongioì LM, La Vignera S, Calogero AE. Molecular biology of spermatogenesis: novel targets of apparently idiopathic male infertility. Int J Mol Sci. 2020;21:E1728.
14. Bourc'his D, Bestor TH. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. Nature. 2004;431:96–9.
15. Barau J, Teissandier A, Zamudio N, Roy S, Nalesso V, Hérault Y, et al. The DNA methyltransferase DNMT3C protects male germ cells from transposon activity. Science. 2016;354:909–12.
16. Song N, Endo D, Song B, Shibata Y, Koji T. 5-aza-2′-deoxycytidine impairs mouse spermatogenesis at multiple stages through different usage of DNA methyltransferases. Toxicology. 2016;361–362:62–72.
17. Lambrot R, Xu C, Saint-Phar S, Chountalos G, Cohen T, Paquet M, et al. Low paternal dietary folate alters the mouse sperm epigenome and is associated with negative pregnancy outcomes. Nat Commun. 2013;4:2889.
18. Radford EJ, Ito M, Shi H, Corish JA, Yamazawa K, Isganaitis E, et al. In utero effects. In utero undernourishment perturbs the adult sperm methylome and intergenerational metabolism. Science. 2014;345:1255903.
19. de Castro BT, Ingerslev LR, Alm PS, Versteyhe S, Massart J, Rasmussen M, et al. High-fat diet reprograms the epigenome of rat spermatozoa and transgenerationally affects metabolism of the offspring. Mol Metab. 2016;5:184–97.
20. Milekic MH, Xin Y, O'Donnell A, Kumar KK, Bradley-Moore M, Malaspina D, et al. Age-related sperm DNA methylation changes are transmitted to offspring and associated with abnormal behavior and dysregulated gene expression. Mol Psychiatry. 2015;20:995–1001.
21. Maurice C, Dalvai M, Lambrot R, Deschênes A, Scott-Boyer M-P, McGraw S, et al. Early-life exposure to environmental contaminants perturbs

the sperm epigenome and induces negative pregnancy outcomes for three generations via the paternal lineage. Epigenomes. 2021;5:10.
22. Abbasi M, Smith AD, Swaminathan H, Sangngern P, Douglas A, Horsager A, et al. Establishing a stable, repeatable platform for measuring changes in sperm DNA methylation. Clin Epigenet. 2018;10:119.
23. Boissonnas CC, Abdalaoui HE, Haelewyn V, Fauque P, Dupont JM, Gut I, et al. Specific epigenetic alterations of IGF2-H19 locus in spermatozoa from infertile men. Eur J Hum Genet EJHG. 2010;18:73–80.
24. Laqqan M, Tierling S, Alkhaled Y, LoPorto C, Hammadeh ME. Alterations in sperm DNA methylation patterns of oligospermic males. Reprod Biol. 2017;17:396–400.
25. Tang Q, Pan F, Yang J, Fu Z, Lu Y, Wu X, et al. Idiopathic male infertility is strongly associated with aberrant DNA methylation of imprinted loci in sperm: a case-control study. Clin Epigenetics. 2018;10:134.
26. Aston KI, Uren PJ, Jenkins TG, Horsager A, Cairns BR, Smith AD, et al. Aberrant sperm DNA methylation predicts male fertility status and embryo quality. Fertil Steril. 2015;104:1388-1397.e5.
27. Urdinguio RG, Bayón GF, Dmitrijeva M, Toraño EG, Bravo C, Fraga MF, et al. Aberrant DNA methylation patterns of spermatozoa in men with unexplained infertility. Hum Reprod Oxf Engl. 2015;30:1014–28.
28. Jenkins TG, Aston KI, Meyer TD, Hotaling JM, Shamsi MB, Johnstone EB, et al. Decreased fecundity and sperm DNA methylation patterns. Fertil Steril. 2016;105:51-57.e1-3.
29. Åsenius F, Danson AF, Marzi SJ. DNA methylation in human sperm: a systematic review. Hum Reprod Update. 2020;26:841–73.
30. Santi D, De Vincentis S, Magnani E, Spaggiari G. Impairment of sperm DNA methylation in male infertility: a meta-analytic study. Andrology. 2017;5:695–703.
31. Leitão E, Di Persio S, Laurentino S, Wöste M, Dugas M, Kliesch S, et al. The sperm epigenome does not display recurrent epimutations in patients with severely impaired spermatogenesis. Clin Epigenet. 2020;12:61.
32. Jenkins TG, Aston KI, James ER, Carrell DT. Sperm epigenetics in the study of male fertility, offspring health, and potential clinical applications. Syst Biol Reprod Med. 2017;63:69–76.
33. Fair S, Lonergan P. Review: understanding the causes of variation in reproductive wastage among bulls. Anim Int J Anim Biosci. 2018;12:s53-62.
34. Özbek M, Hitit M, Kaya A, Jousan FD, Memili E. Sperm functional genome associated with bull fertility. Front Vet Sci. 2021;8:571.
35. Kropp J, Carrillo JA, Namous H, Daniels A, Salih SM, Song J, et al. Male fertility status is associated with DNA methylation signatures in sperm and transcriptomic profiles of bovine preimplantation embryos. BMC Genomics. 2017;18:280.
36. Gross N, Peñagaricano F, Khatib H. Integration of whole-genome DNA methylation data with RNA sequencing data to identify markers for bull fertility. Anim Genet. 2020;51:502–10.
37. Verma A, Rajput S, De S, Kumar R, Chakravarty AK, Datta TK. Genome-wide profiling of sperm DNA methylation in relation to buffalo (Bubalus bubalis) bull fertility. Theriogenology. 2014;82:750-759.e1.
38. Fang L, Zhou Y, Liu S, Jiang J, Bickhart DM, Null DJ, et al. Comparative analyses of sperm DNA methylomes among human, mouse and cattle provide insights into epigenomic evolution and complex traits. Epigenetics. 2019;14:260–76.
39. Takeda K, Kobayashi E, Ogata K, Imai A, Sato S, Adachi H, et al. Differentially methylated CpG sites related to fertility in Japanese Black bull spermatozoa: epigenetic biomarker candidates to predict sire conception rate. J Reprod Dev. 2021.
40. Narud B, Khezri A, Zeremichael TT, Stenseth E-B, Heringstad B, Johannisson A, et al. Sperm chromatin integrity and DNA methylation in Norwegian Red bulls of contrasting fertility. Mol Reprod Dev. 2021;88:187–200.
41. Takeda K, Kobayashi E, Nishino K, Imai A, Adachi H, Hoshino Y, et al. Age-related changes in DNA methylation levels at CpG sites in bull spermatozoa and in vitro fertilization-derived blastocyst-stage embryos revealed by combined bisulfite restriction analysis. J Reprod Dev. 2019;65:305–12.
42. Lambert S, Blondin P, Vigneault C, Labrecque R, Dufort I, Sirard M-A. Spermatozoa DNA methylation patterns differ due to peripubertal age in bulls. Theriogenology. 2018;106:21–9.
43. Perrier J-P, Sellem E, Prézelin A, Gasselin M, Jouneau L, Piumi F, et al. A multi-scale analysis of bull sperm methylome revealed both species peculiarities and conserved tissue-specific features. BMC Genomics. 2018;19:1–18.
44. Jiang Z, Lin J, Dong H, Zheng X, Marjani SL, Duan J, et al. DNA methylomes of bovine gametes and in vivo produced preimplantation embryos. Biol Reprod. 2018;99:949–59.
45. El Hajj N, Zechner U, Schneider E, Tresch A, Gromoll J, Hahn T, et al. Methylation status of imprinted genes and repetitive elements in sperm DNA from infertile males. Sex Dev Genet Mol Biol Evol Endocrinol Embryol Pathol Sex Determ Differ. 2011;5:60–9.
46. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6:11.
47. Zhou Y, Connor EE, Bickhart DM, Li C, Baldwin RL, Schroeder SG, et al. Comparative whole genome DNA methylation profiling of cattle sperm and somatic tissues reveals striking hypomethylated patterns in sperm. GigaScience. 2018;7:giy039.
48. Feng H, Conneely KN, Wu H. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. Nucleic Acids Res. 2014;42:e69.
49. Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol. 2012;13:R87.
50. Taudt A, Colomé-Tatché M, Johannes F. Genetic sources of population epigenomic variation. Nat Rev Genet. 2016;17:319–32.
51. Jongbloets BC, Pasterkamp RJ. Semaphorin signalling during development. Dev Camb Engl. 2014;141:3292–7.
52. Camprubí C, Salas-Huetos A, Aiese-Cigliano R, Godo A, Pons M-C, Castellano G, et al. Spermatozoa from infertile patients exhibit differences of DNA methylation associated with spermatogenesis-related processes: an array-based analysis. Reprod Biomed Online. 2016;33:709–19.
53. Seiler Vellame D, Castanho I, Dahir A, Mill J, Hannon E. Characterizing the properties of bisulfite sequencing data: maximizing power and sensitivity to identify between-group differences in DNA methylation. BMC Genomics. 2021;22:446.
54. Zhang W, Spector TD, Deloukas P, Bell JT, Engelhardt BE. Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. Genome Biol. 2015;16:14.
55. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, et al. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biol. 2011;12:R10.
56. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, et al. DNA methylation profiling of human chromosomes 6, 20 and 22. Nat Genet. 2006;38:1378–85.
57. Paun O, Verhoeven KJF, Richards CL. Opportunities and limitations of reduced representation bisulfite sequencing in plant ecological epigenomics. New Phytol. 2019;221:738–42.
58. Qi Y. Random forest for bioinformatics. Ensemble machine learning: methods and applications. New York: Springer; 2012.
59. Donkin I, Barrès R. Sperm epigenetics and influence of environmental factors. Mol Metab. 2018;14:1–11.
60. Butcher LM, Beck S. AutoMeDIP-seq: a high-throughput, whole genome, DNA methylation assay. Methods San Diego Calif. 2010;52:223–31.
61. Czyz W, Morahan JM, Ebers GC, Ramagopalan SV. Genetic, environmental and stochastic factors in monozygotic twin discordance with a focus on epigenetic differences. BMC Med. 2012;10:93.
62. Jenkins TG, Aston KI, Carrell DT. Sperm epigenetics and aging. Transl Androl Urol. 2018;7:S328–35.
63. Xie K, Ryan DP, Pearson BL, Henzel KS, Neff F, Vidal RO, et al. Epigenetic alterations in longevity regulators, reduced life span, and exacerbated aging-related pathology in old father offspring mice. Proc Natl Acad Sci U S A. 2018;115:E2348–57.
64. Kiefer H, Perrier J-P. DNA methylation in bull spermatozoa: evolutionary impacts, interindividual variability, and contribution to the embryo. Can J Anim Sci. 2019;100:1–16.
65. Perrier J-P, Kenny DA, Chaulot-Talmon A, Byrne CJ, Sellem E, Jouneau L, et al. Accelerating onset of puberty through modification of early life nutrition induces modest but persistent changes in bull sperm DNA methylation profiles post-puberty. Front Genet. 2020;11:945.
66. Rahman MB, Schellander K, Luceño NL, Van Soom A. Heat stress responses in spermatozoa: mechanisms and consequences for cattle fertility. Theriogenology. 2018;113:102–12.

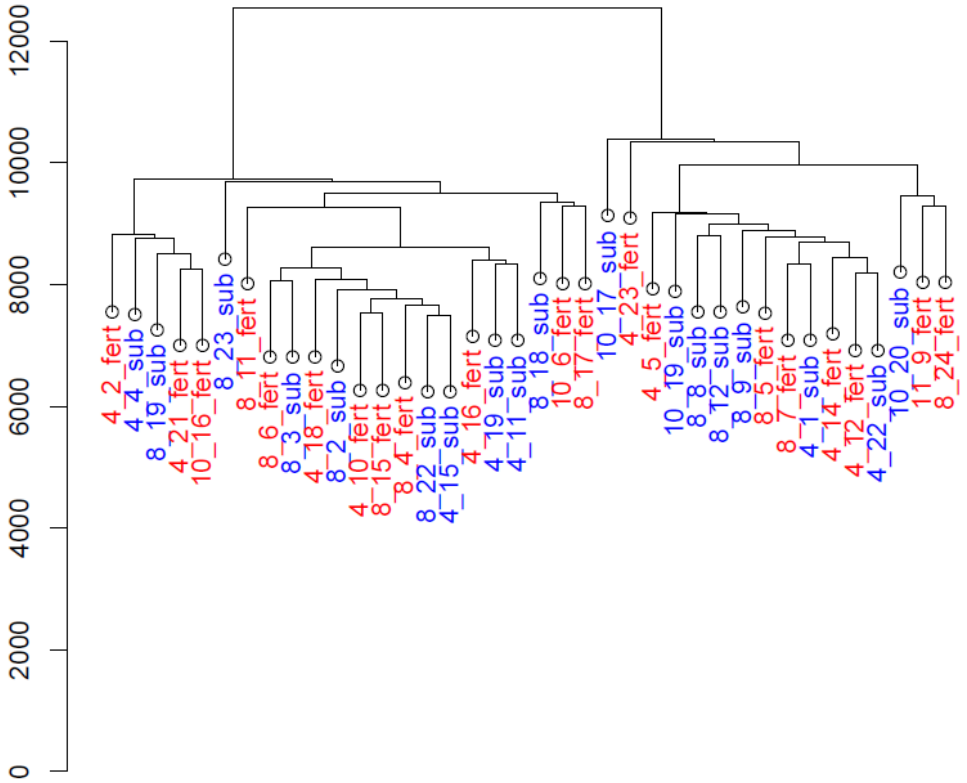


Figure 28 : Les CpG10 ne permettent pas de différencier les animaux fertiles des animaux subfertiles en race Holstein. Une classification hiérarchique a été réalisée en se basant sur l'information de méthylation. Les labels ont été coloriés en fonction de la fertilité des animaux.

67. Wang H, Wan H, Li X, Liu W, Chen Q, Wang Y, et al. Atg7 is required for acrosome biogenesis during spermatogenesis in mice. Cell Res. 2014;24:852–69.
68. Shang Y, Wang H, Jia P, Zhao H, Liu C, Liu W, et al. Autophagy regulates spermatid differentiation via degradation of PDLIM1. Autophagy. 2016;12:1575–92.
69. Long SKR, Fulkerson E, Breese R, Hernandez G, Davis C, Melton MA, et al. A comparison of midline and tracheal gene regulation during Drosophila development. PLoS ONE. 2014;9:e85518.
70. Jiang J, Ma L, Prakapenka D, VanRaden PM, Cole JB, Da Y. A large-scale genome-wide association study in U.S. Holstein Cattle. Front Genet. 2019;10:412.
71. Boggild S, Molgaard S, Glerup S, Nyengaard JR. Highly segregated localization of the functionally related vps10p receptors sortilin and SorCS2 during neurodevelopment. J Comp Neurol. 2018;526:1267–86.
72. Rezgaoui M, Hermey G, Riedel IB, Hampe W, Schaller HC, Hermans-Borgmeyer I. Identification of SorCS2, a novel member of the VPS10 domain containing receptor family, prominently expressed in the developing mouse brain. Mech Dev. 2001;100:335–8.
73. Krüger M, Schäfer K, Braun T. The homeobox containing gene Lbx1 is required for correct dorsal–ventral patterning of the neural tube. J Neurochem. 2002;82:774–82.
74. Sabo MC, Nath K, Elinson RP. Lbx1 expression and frog limb development. Dev Genes Evol. 2009;219:609–12.
75. Konstanze S, Petra N, Julia K, Thomas B. The homeobox gene Lbx1 specifies a subpopulation of cardiac neural crest necessary for normal heart development. Circ Res. 2003;92:73–80.
76. De Graeve F, Jagla T, Daponte J-P, Rickert C, Dastugue B, Urban J, et al. The ladybird homeobox genes are essential for the specification of a subpopulation of neural cells. Dev Biol. 2004;270:122–34.
77. Deng S, Hirschberg A, Worzfeld T, Penachioni JY, Korostylev A, Swiercz JM, et al. Plexin-B2, but not Plexin-B1, critically modulates neuronal migration and patterning of the developing nervous system in vivo. J Neurosci Off J Soc Neurosci. 2007;27:6333–47.
78. Worzfeld T, Offermanns S. Semaphorins and plexins as therapeutic targets. Nat Rev Drug Discov. 2014;13:603–21.
79. Sujit KM, Sarkar S, Singh V, Pandey R, Agrawal NK, Trivedi S, et al. Genome-wide differential methylation analyses identifies methylation signatures of male infertility. Hum Reprod Oxf Engl. 2018;33:2256–67.
80. Ferreira AO, Vasconcelos BG, Favaron PO, Santos AC, Leandro RM, Pereira FTV, et al. Bovine central nervous system development. Pesqui Veterinária Bras. 2018;38:147–53.
81. Pértille F, Alvarez-Rodriguez M, da Silva AN, Barranco I, Roca J, Guerrero-Bosagna C, et al. Sperm methylome profiling can discern fertility levels in the porcine biomedical model. Int J Mol Sci. 2021;22:2679.
82. Houshdaran S, Cortessis VK, Siegmund K, Yang A, Laird PW, Sokol RZ. Widespread epigenetic abnormalities suggest a broad DNA methylation erasure defect in abnormal human sperm. PLoS ONE. 2007;2:e1289.
83. Qu J, Hodges E, Molaro A, Gagneux P, Dean MD, Hannon GJ, et al. Evolutionary expansion of DNA hypomethylation in the mammalian germline genome. Genome Res. 2018;28:145–58.
84. Schübeler D. Function and information content of DNA methylation. Nature. 2015;517:321–6.
85. Adelson DL, Raison JM, Edgar RC. Characterization and distribution of retrotransposons and simple sequence repeats in the bovine genome. Proc Natl Acad Sci U S A. 2009;106:12855–60.
86. Newkirk SJ, Lee S, Grandi FC, Gaysinskaya V, Rosser JM, Berg NV, et al. Intact piRNA pathway prevents L1 mobilization in male meiosis. Proc Natl Acad Sci. 2017;114:E5635–44.
87. Heyn H, Ferreira HJ, Bassas L, Bonache S, Sayols S, Sandoval J, et al. Epigenetic disruption of the PIWI pathway in human spermatogenic disorders. PLoS ONE. 2012;7:e47892.
88. Cheng Y-S, Wee S-K, Lin T-Y, Lin Y-M. MAEL promoter hypermethylation is associated with de-repression of LINE-1 in human hypospermatogenesis. Hum Reprod Oxf Engl. 2017;32:2373–81.
89. Jachowicz JW, Bing X, Pontabry J, Bošković A, Rando OJ, Torres-Padilla M-E. LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. Nat Genet. 2017;49:1502–10.
90. Lou C, Goodier JL, Qiang R. A potential new mechanism for pregnancy loss: considering the role of LINE-1 retrotransposons in early spontaneous miscarriage. Reprod Biol Endocrinol RBE. 2020;18:6.
91. Barbat A, Le Mezec P, Ducrocq V, Mattalia S, Fritz S, Boichard D, et al. Female fertility in french dairy breeds: current situation and strategies for improvement. J Reprod Dev. 2010;56:S15-21.
92. Sellem E, Broekhuijse MLWJ, Chevrier L, Camugli S, Schmitt E, Schibler L, et al. Use of combinations of in vitro quality assessments to predict fertility of bovine semen. Theriogenology. 2015;84:1447-1454.e5.
93. Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. Nat Protoc. 2011;6:468–81.
94. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinf Oxf Engl. 2011;27:1571–2.
95. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25.
96. Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brøndum RF, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet. 2014;46:858–65.
97. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29:24–6.
98. Le S, Josse J, Husson F. FactoMineR: an R package for multivariate analysis. J Stat Softw. 2008;25:1–18.
99. Ignatiadis N, Klaus B, Zaugg JB, Huber W. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. Nat Methods. 2016;13:577–80.
100. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4:44–57.
101. Josse J, Husson F. missMDA: a package for handling missing values in multivariate data analysis. J Stat Softw. 2016;70:1–31.
102. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
103. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006;27:861–74.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC

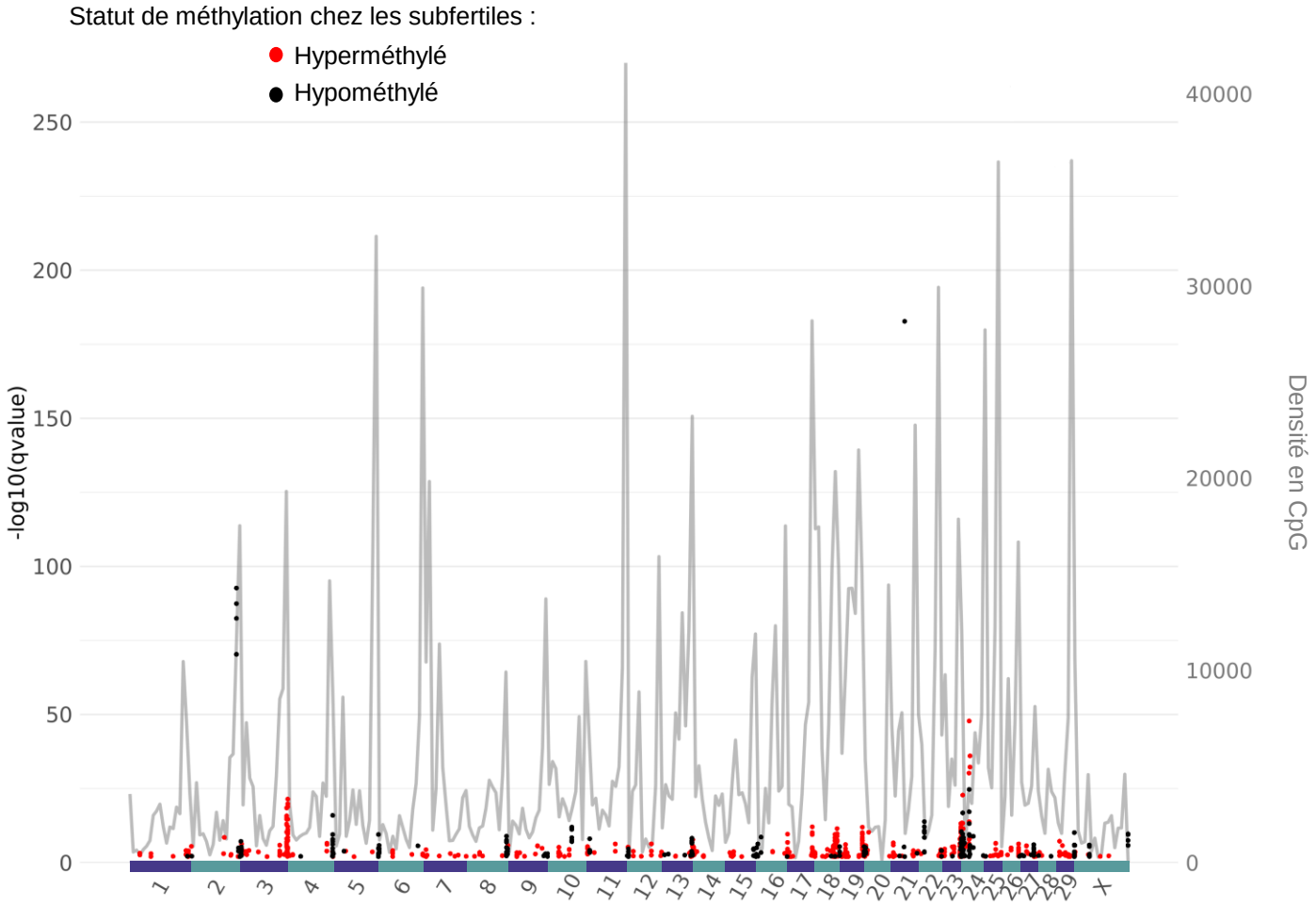


Figure 29 : Les DMC sont réparties sur tous les chromosomes. Chaque point représente une DMC et est colorié en rouge si la DMC est hyperméthylée chez les taureaux subfertiles ou en noir si elle est hypométhylée. La significativité des DMC est indiquée par l'axe y gauche décrivant le logarithme de la qvalue des analyses différentielles. La courbe en gris clair représente la densité des CpG analysés ; l'échelle est indiquée sur l'axe y droit. L'axe x lui, représente les chromosomes.

I.II.III : Analyse en race Holstein

Dispositif expérimental

De la même façon qu'en race Montbéliarde, une analyse différentielle a été réalisée dans la race Holstein pour identifier des DMC liées à la fertilité. Pour cela les animaux extrêmes de la cohorte « fertilité » en Holstein ont été utilisés (Figure 20 et Annexe 4), soit 38 animaux dont 20 ont été classifiés comme fertiles et 18 comme subfertiles. L'extraction de l'ADN spermatique, les techniques expérimentales mettant en évidence la méthylation de l'ADN et le pipeline bio-informatique sont exactement les mêmes que dans l'étude précédente et ne seront donc pas re-détaillés ici. La différence par rapport au travail précédent est l'absence de cohorte de validation indépendante dans le but de valider les résultats obtenus sur la cohorte principale.

La qualité des séquences et les résultats de l'alignement sont présentés en Annexe 5. On peut observer l'absence de différences pour ces paramètres entre les animaux fertiles et subfertiles, excluant de potentiels biais dus à des différences de qualité sur les résultats obtenus.

Analyses descriptive et différentielle

Dans le but de déterminer si des différences globales au niveau du méthylome spermatique pouvaient être liées à la fertilité, une classification hiérarchique a été réalisée en se basant sur l'information de méthylation des 1 634 038 CpG du background (CpG10) obtenus après suppression des potentiels CpG polymorphes (Figure 28). Cette classification ne regroupe pas les animaux en fonction de leur classe de fertilité, suggérant que d'éventuelles différences du méthylome spermatique en fonction de la fertilité n'affectent pas la globalité des CpG analysés. Une analyse différentielle de méthylation a ensuite été conduite CpG par CpG. Cette analyse a été réalisée à l'aide du logiciel methylKit entre les animaux fertiles et subfertiles, sur les 1 634 038 CpG10, dans les conditions décrites dans l'article 1 (Costes *et al.*, 2022). Ainsi, 738 DMC ont été mises en évidence, parmi lesquelles 533 sont hyperméthylées chez les animaux subfertiles et 235 sont hypométhylées chez les animaux subfertiles.

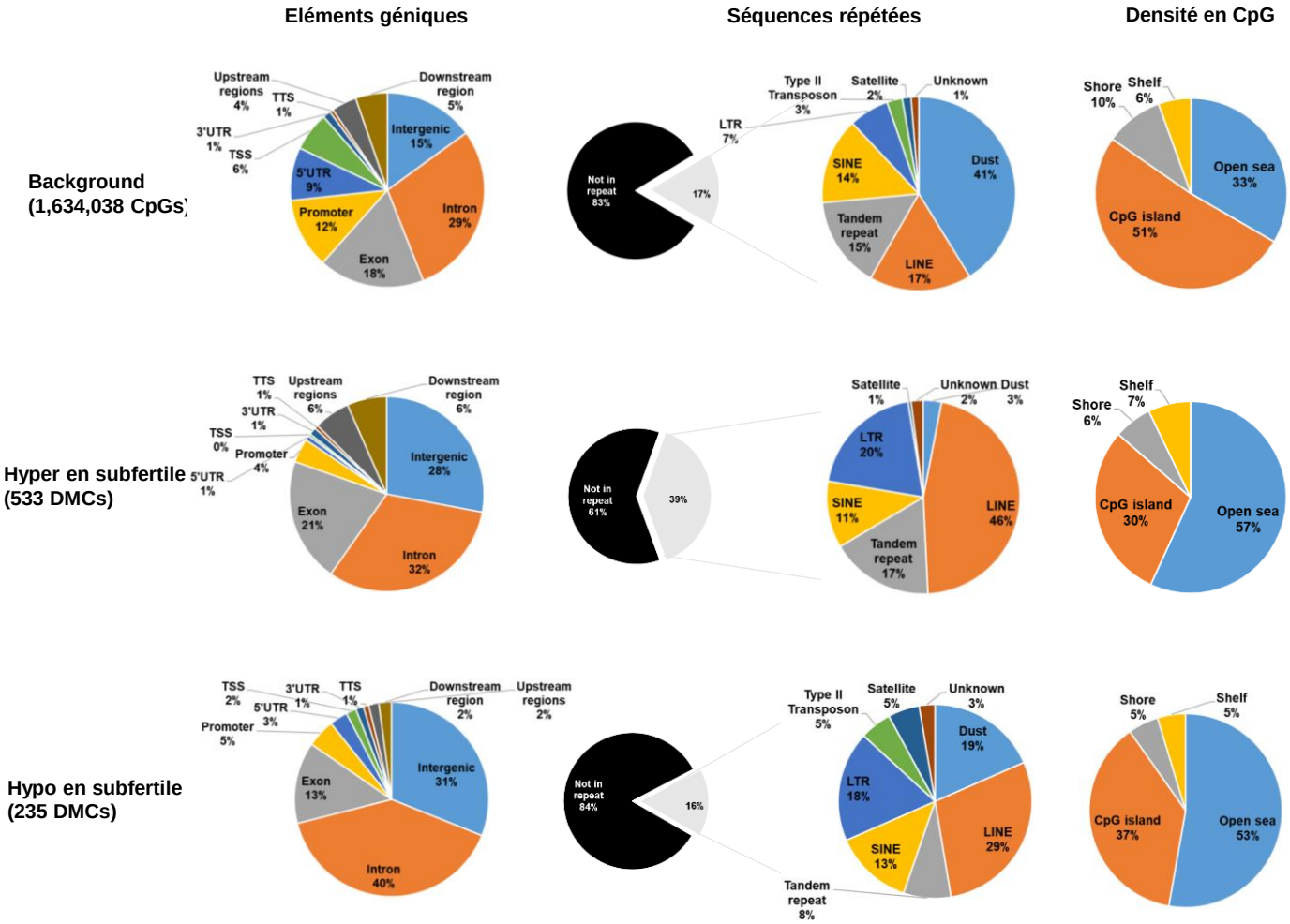


Figure 30 : Localisation génomique des DMC. Les CpG du background et les DMC ont été annotés comme décrit dans la partie « Method » de l'article 1. La proportion de chaque type d'annotation au sein de trois groupes fonctionnels (éléments géniques, séquences répétées et densité en CpG) est représentée pour le background et les DMC hyper ou hypométhylées chez les taureaux subfertiles.

La position chromosomique des DMC en fonction des résultats de l'analyse différentielle est présentée Figure 29. Les DMC sont présentes sur l'intégralité des chromosomes et semblent être regroupées par statut de méthylation.

Ce résultat suggère qu'il existe des différences dans le méthylome spermatique de taureaux Holstein ayant des fertilités contrastées.

Localisation des DMC dans les éléments fonctionnels du génome

Dans un second temps, nous nous sommes intéressés aux régions du génome enrichies en DMC par comparaison avec le background de l'analyse. Pour cela, les CpG appartenant au background et les DMC hypo ou hyperméthylées chez les subfertiles, ont été annotées en fonction des éléments géniques, des séquences répétées et de la densité en CpG. Les proportions obtenues pour chaque élément géniques sont représentées Figure 30.

Comparé au background, les DMC hyperméthylées sont enrichies au niveau des séquences répétées (17% des CpG sont dans des séquences répétées dans le background comparé à 39% dans les DMC hyperméthylées). De plus, au sein des séquences répétées on observe un enrichissement des séquences touchant les LINE (17% dans le background contre 46% dans les DMC hyperméthylées chez les subfertiles), et dans les LTR (7% dans le background contre 20% dans les DMC hyperméthylées chez les subfertiles).

Les DMC hypométhylées elles, ne sont pas enrichies au niveau des séquences répétées (17% dans le background et 16% dans les DMC hypométhylées). Néanmoins, la représentation des différentes catégories d'annotations entre le background et ces DMC varient. En effet tout comme les DMC hyperméthylées chez les animaux fertiles, on observe une augmentation de la proportion des LINE (29%) et des LTR (18%) dans les DMC hypométhylées chez les taureaux subfertiles.

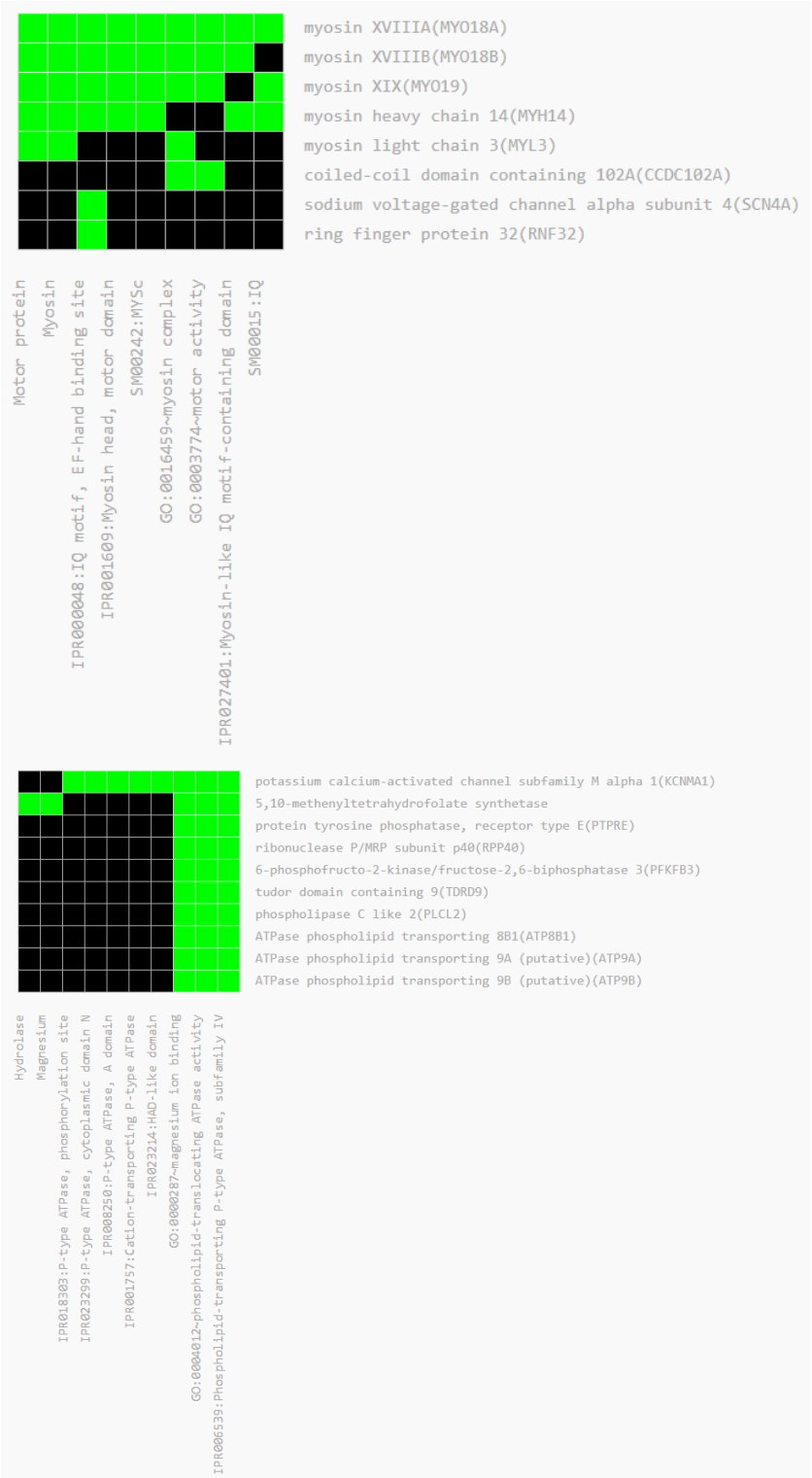


Figure 31 : Résultats des analyses d'enrichissement par DAVID. Analyse d'enrichissement fonctionnel réalisée parDAVID en soumettant les gènes contenant des DMC contre les gènes contenant des CpG du background. Seuls 140 gènes différentiellement méthylés ont été pris en compte par DAVID. Les deux groupes de termes significatifs (EASE score > 1.3) ont été représentés dans cette figure. La correspondance entre un terme et un gène est représentée en vert.

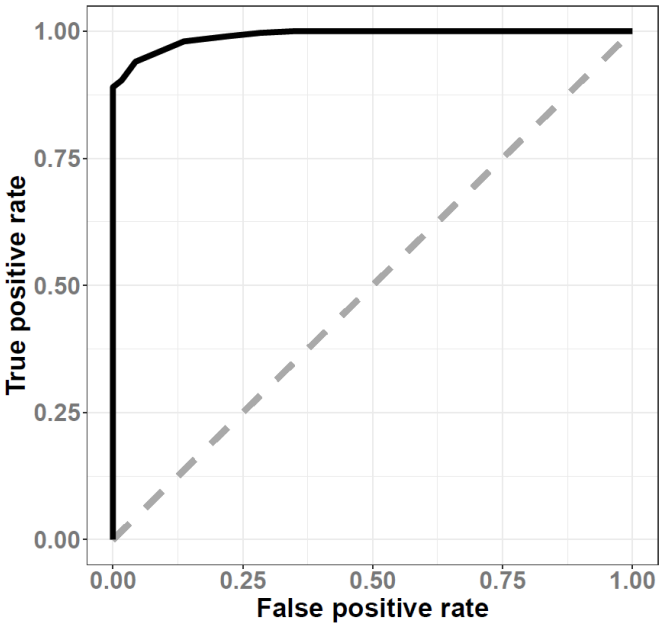
Comme expliqué dans les résultats obtenus en race Montbéliarde, le contrôle de ces séquences par la méthylation de l'ADN est critique pour la fertilité. Il est donc intéressant de constater un déséquilibre de représentation de ces séquences entre le background et les DMC.

Gènes différentiellement méthylés

Une analyse des gènes contenant des DMC a également été réalisée. Pour cela une analyse d'enrichissement fonctionnel a été effectuée avec la base de données DAVID (« Database for Annotation, Visualization and Integrated Discovery » (Huang *et al.,* 2009)) en utilisant le background comme référence, qui ne représente pas tous les gènes. Deux groupes de termes significativement enrichis (EASE score > 1,3) ont pu être mis en évidence et sont présentés figure 31.

Le premier groupe de termes est majoritairement composé de gènes appartenant à la famille des myosines. Au sein de ce groupe on remarque 3 gènes ayant des fonctions intéressantes. Tout d'abord les gènes codant les deux myosines de type 18 (A et B). Elles sont connues comme étant importantes pour la fonction des myocytes cardiaques au cours du développement. De plus, l'inactivation de ces gènes de manière individuelle est létale au stade embryonnaire à l'état homozygote (Ajima *et al.*, 2008; Horsthemke *et al.*, 2019). En s'intéressant aux termes GO associés aux gènes de ce groupe, on remarque que le terme « IQ motif » est souvent présent et est partagé par beaucoup de gènes. Les motifs IQ sont des motifs protéiques permettant l'interaction avec les calmodulines (Bähler and Rhoads, 2002). Les protéines présentant ces motifs ont des fonctions associées avec la division cellulaire, l'organisation du cytosquelette et la fécondation.

Dans le second groupe de termes, on peut remarquer la présence importante de gènes impliqués dans les « Phospholipid transporting ATPases ». Je n'ai pour l'instant pas trouvé de lien direct entre cette famille de gènes et la fertilité mâle. Néanmoins, parmi les gènes impliqués dans ce groupe on peut observer la présence de *TDRD9*. Il est impliqué dans la fertilité mâle, en particulier au moment du développement des cellules germinales mâles afin de réguler la méthylation de l'ADN des séquences répétées de type LINE L1 (Shoji *et al.*, 2009). Son inactivation à l'état homozygote a pour conséquence



Précision	AUC	Sensibilité	Spécificité
0.94	0.99	0.96	0.91

Figure 32. Capacité de prédiction du modèle élaboré à partir des DMC spermatiques dans la race Holstein. A gauche se trouve la courbe ROC moyenne obtenue à partir des 50 jeux d'apprentissage et de test. A droite se trouve une table contenant la moyenne des performances obtenue sur l'échantillon de test après 50 itérations, pour la précision du modèle, l'AUC, la sensibilité et la spécificité

une infertilité mâle avec un blocage méiotique. Les DMC sont localisées dans un intron. Il est peu probable que l'expression de ce gène ait été drastiquement modifiée par la méthylation différentielle, ce qui aurait sans doute eu des conséquences plus importantes sur la fertilité.

Ces résultats montrent qu'une partie des gènes contenant des DMC ont des fonctions potentiellement en lien avec la fertilité. Pour compléter ces résultats, il serait également intéressant de réaliser une étude bibliographique systématique sur l'ensemble des gènes différentiellement méthylés comme en race Montbéliarde.

Modèle de prédiction

Un autre aspect de ce travail a consisté à construire un modèle permettant de prédire la classe de fertilité des taureaux. Pour réaliser cela, les individus ont été répartis entre un groupe d'entraînement (2/3 des animaux) et un groupe de test (le tiers restant). Le modèle est construit sur le jeu d'entrainement, et il est testé (comparaison entre fertilité prédite et fertilité réelle) sur le jeu de test. A partir des prédictions, quatre indicateurs de la qualité du modèle sont calculés : la précision, l'AUC, la sensibilité et la spécificité du modèle. Afin d'avoir une vue d'ensemble qui ne soit pas dépendante d'une seule combinaison entre individus, 50 ré-échantillonnages aléatoires sont effectués entre le jeu d'entraînement et le jeu de test, en conservant la proportion originale d'animaux fertiles et subfertiles dans chaque jeu d'entrainement et jeu de test. Les indicateurs fournis correspondent alors à la moyenne des indicateurs obtenus par les 50 itérations. Les résultats sont présentés dans la Figure 32.

On remarque que les performances du modèle sont globalement excellentes, avec une précision de 0.94 et une AUC de 0.99. Néanmoins, un des défauts de la cohorte utilisée en race Holstein est sa faible taille, ce qui n'amène qu'à 12 animaux dans la cohorte de test. Ainsi, les résultats obtenus sont certes intéressants, mais à modérer en raison des effectifs réduits. Afin de confirmer les résultats obtenus dans ce travail, il faudra réaliser des prédictions sur des cohortes de taureaux Holstein indépendantes, qui sont en cours de constitution.

I.II.IV : Comparaison des résultats obtenus dans les deux races

Taille intersection	0	1	2	3	4	5
Fréquence	40262	8752	910	73	2	1

Figure 33 : L'intersection entre les DMC des deux races n'est pas liée au hasard. Une simulation a été réalisée dans laquelle : 738 DMC en Holstein ont été tirées au hasard parmi le background de la race Holstein ; 492 DMC ont été tirées au hasard parmi le background de la race Montbéliarde ; l'intersection des DMC entre ces deux tirages aléatoires a ensuite été examinée. Cette expérience a été réalisée 50 000 fois et les résultats sont présentés dans le tableau, où sont indiqués la taille de l'intersection (en nombre de DMC) et le nombre de fois où cette taille a été atteinte (fréquence).

Les deux études réalisées ont chacune eu pour objectif d'analyser le méthylome spermatique des taureaux en relation avec la fertilité mâle. Les taureaux des deux races ont été élevés dans des régions différentes, avant de rejoindre des centres de production avec une implantation locale différente et qui dépendent d'entreprises de sélection différentes (Evolution pour les taureaux Holstein, avec une implantation dans l'Ouest de la France, et Umotest et Evajura pour les taureaux de race Montbéliarde, localisés dans l'Est dans des zones de moyennes montagnes). Les taureaux des deux races présentent également des différences génétiques et sont soumis à des pratiques d'élevage différentes, renforçant l'intérêt de comparer les résultats obtenus dans chacune des races.

Dans un premier temps, on peut remarquer que les groupes de fertilité ne peuvent pas être discriminés sur la base de l'ensemble des CpG10, et ce dans les deux races. Cela n'est pas illogique quand on prend en compte les résultats de l'analyse différentielle. En effet, dans les deux races seules quelques centaines de DMC ont pu être identifiées à partir de plus d'1,5 millions de CpG (background), ce qui représente moins de 0,03 % des CpG10 qui varient en fonction de la fertilité des animaux.

Dans les deux races on observe également une tendance similaire au niveau du statut de méthylation des DMC. En effet, les DMC présentant une hyperméthylation chez les animaux subfertiles sont majoritaires.

Entre ces deux études, le nombre de DMC en commun s'élève à 14. Cela peut sembler faible car représentant uniquement 1,1% du total des DMC analysés dans les deux races (738 en Holstein et 492 en Montbéliard). Cependant, les backgrounds au sein des deux races représentent chacun plus d'1,5 million de CpG. Ainsi, avoir des DMC communs entre races est un évènement peu probable. Pour illustrer ce phénomène, nous avons simulé un tirage aléatoire de DMC au sein des backgrounds des deux races respectives, afin de regarder les éléments communs entre les races (le détail se trouve dans la légende de la figure 33). On peut remarquer que le maximum d'évènements communs observés au cours de 50 000 simulations est de 5 et que cet évènement n'a été obtenu qu'une fois sur 50 000. Ainsi, obtenir 14 évènements communs est hautement improbable. On peut donc supposer que malgré le

faible nombre de DMC communs entre races, ce dernier a de faibles chances d'être lié au hasard, suggérant une potentielle signification biologique pour les DMC en commun.

Un aspect intéressant partagé par les deux races sont les éléments génomiques touchés par les DMC. En effet, dans les deux races on observe un enrichissement des DMC dans les éléments intergéniques aux dépends d'éléments géniques et en particulier des promoteurs, sites d'initiation de la transcription et des régions non traduites (Figure 30 et Figure 4 de l'article 1). On observe également une différence : dans la race Montbéliarde ce sont les DMC hypométhylées chez les taureaux subfertiles qui sont enrichies en régions répétées alors que dans la race Holstein, ce sont les DMC hyperméthylées chez les taureaux subfertiles qui le sont. Néanmoins, malgré cette différence on peut remarquer que les éléments répétés enrichis sont les mêmes : ce sont les LINE et les LTR. Comme discuté dans l'article en race Montbéliarde, le contrôle de ces régions est essentiel au cours de la différenciation des cellules germinales mâles et au cours du développement embryonnaire. Ainsi, il est intéressant d'observer un enrichissement de ces régions dans deux races.

En s'intéressant aux résultats des analyses d'enrichissement, on remarque des différences entre les races. Dans la race Holstein, les groupes de termes sont en lien avec le transfert de lipides, le transport ionique et les gènes codant pour des protéines à motif IQ. Dans la race Montbéliarde en revanche, le seul groupe de termes significativement enrichis était majoritairement porté par des gènes codant des plexines, qui ont un rôle dans le développement du système nerveux. Dans l'hypothèse où les différences de fertilité entre taureaux auraient pour cause la méthylation de l'ADN, des mécanismes spécifiques à chaque race pourraient ainsi être mis en jeu. Néanmoins, en comparant les gènes différentiellement méthylés dans les deux races, une intersection de 20 gènes a pu être identifiée, suggérant qu'une partie des fonctions touchées est conservée. De manière intéressante, parmi ces 20 gènes, 8 sont impliqués dans des processus de développement *(CDK5RAP2, FHOD3, LRFN3, LRRC55, MBOAT7, PADI2, PLXNA1 et TMEM119)*. Ainsi, malgré l'absence de groupes de termes communs

présentant un enrichissement fonctionnel, il a été possible d'identifier des gènes différentiellement méthylés en commun partageant des fonctions qui peuvent être liées à la fertilité mâle.

Enfin en s'intéressant aux capacités de prédiction, on remarque que les performances sont contrastées entre les races. En effet, dans la race Montbéliarde on observe des performances plus faibles qu'en race Holstein. Néanmoins dans les deux races on remarque que les modèles prédisent globalement bien la classe de fertilité des animaux.

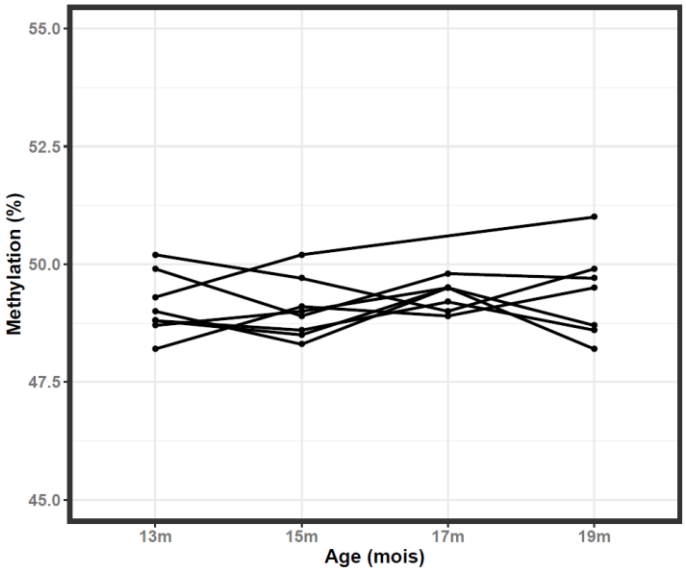
I.III : Etudes longitudinales : cohortes « âge » et « déviation »

I.III.I : Contexte

Enfin, le dernier travail portant exclusivement sur l'analyse du méthylome spermatique a été l'analyse des cohortes « âge » et « déviation ». L'analyse de ces cohortes répond à deux objectifs différents, mais tous deux importants dans le cadre de l'analyse du méthylome spermatique.

La première cohorte est celle du dispositif « âge ». En effet, il a été montré récemment par deux études chez le bovin que le profil de méthylation spermatique évoluait au cours de la vie de l'animal, avec une hyperméthylation progressive de certains CpG (Takeda *et al.,* 2019) et des modifications d'ampleur importante mais limitées au stade péripubertaire (Lambert *et al.,* 2018). Le méthylome spermatique des taureaux présente donc une certaine dynamique sans que le lien avec la fertilité soit précisément décrit. Cet aspect pourrait altérer les performances des modèles construits pour prédire la fertilité à partir du méthylome et biaiser certains résultats. C'est pourquoi le méthylome spermatique de 8 taureaux de race Holstein a été analysé entre 13 et 19 mois d'âge avec un pas de deux mois entre 2 échantillons successifs (Figure 19). Chaque échantillon a été analysé par RRBS, avec le même protocole expérimental que les analyses précédentes. Pour des raisons techniques liées à la préparation des banques RRBS, deux animaux ne sont pas représentés à tous les âges (un manque à 17 mois et l'autre manque à 19 mois).

A



B

Comparaison	13 vs 19	15 vs 19	17 vs 19
DMC (nb)	14	5	29

C

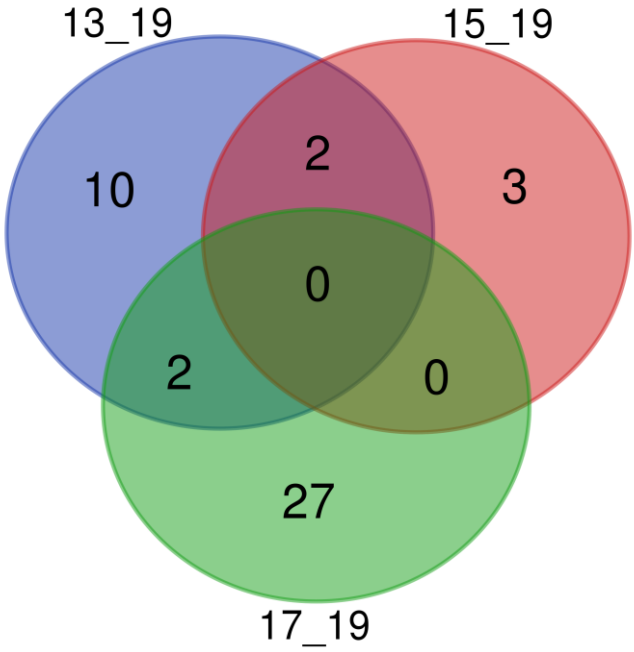


Figure 34 : Etude de la méthylation de l'ADN en fonction de l'âge des animaux. A : Niveau de méthylation moyen (calculé sur l'ensemble des CpG du background), pour chaque animal à chaque âge. B : Table référençant le nombre de DMC pour chaque analyse différentielle. Les analyses différentielles ont été réalisées en utilisant methylKit avec un seuil de différence de méthylation à 10% et un seuil de qvalue de 0, 01. C : Diagramme de Venn illustrant les DMC communes entre les comparaisons montrées en B.

La cohorte « déviation » a également été analysée, afin d'observer si une diminution rapide de fertilité pour des taureaux était liée à une modification du méthylome spermatique. Il est important de noter que les taureaux représentés dans les cohortes « âge » et « déviation » sont également représentés dans la cohorte « fertilité ».

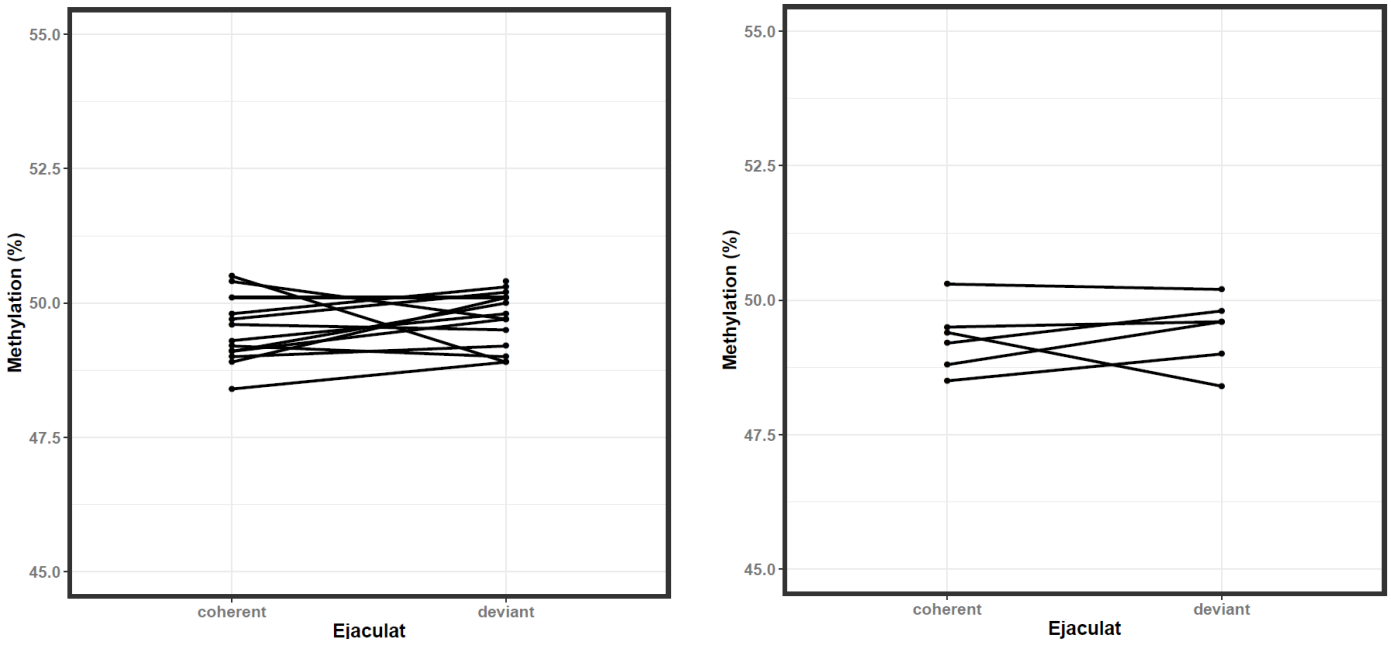
Dans ces deux analyses longitudinales, des échantillons de semence provenant des mêmes animaux ont été comparés. Ainsi, les CpG polymorphes influençant la détection des DMC entre groupes de fertilité ne devraient théoriquement pas avoir d'incidence dans le cas présent. Les informations acquises sur l'ensemble des CpG suffisamment couverts, y compris les CpG polymorphes, auraient donc pu être utilisées. Néanmoins, pour 2 animaux appartenant à la cohorte « âge », le RRBS n'a pas pu être réalisé pour des raisons techniques. Les CpG polymorphes ont donc été supprimés pour éliminer les biais génétiques liés au fait que les groupes d'âge à comparer incluent des animaux différents. En revanche, tous les animaux de la cohorte « déviation » ont été analysés, ce qui a permis de conserver les CpG polymorphes.

I.III.II : Analyse de la cohorte Age

<u>Méthylation moyenne entre 13 et 19 mois</u>

Nous nous sommes donc intéressés à la modification du méthylome spermatique en fonction de l'âge du taureau. Dans un premier temps, nous avons analysé le taux de méthylation moyen sur l'ensemble des CpG10 en fonction de l'âge (Figure 34A). La valeur moyenne de méthylation est comprise entre 48,2% et 51% quelles que soient les conditions, ce qui témoigne d'une variation faible de la méthylation globale au cours de ces âges. Il n'y a pas de variation significative du taux de méthylation en fonction de l'âge des taureaux. On peut tout de même observer qu'un des taureaux semble avoir une augmentation du taux de méthylation entre 13 et 19 mois. Néanmoins, pour des raisons techniques cet animal n'a pas été analysé à 17 mois, ce qui réduit la confiance que l'on peut avoir en ces résultats.

A



B

Comparaison	Cohérent fertile vs déviant	Cohérent subfertiles vs déviant
DMC (nb)	36	13

Figure 35 : A : Niveau de méthylation sur l'ensemble des CpG du background, en fonction de la cohérence par rapport à la fertilité moyenne des taureaux. A gauche, il s'agit des 12 taureaux fertiles et à droite des 6 animaux subfertiles. B : Table référençant le nombre de DMC obtenues après analyse différentielle entre échantillons déviant et cohérent.

On peut donc conclure qu'à un niveau global, le taux de méthylation moyen des CpG10 ne varie pas entre 13 et 19 mois, ce qui n'exclut pas des variations sur des loci CpG individuels.

Analyses différentielles de méthylation

Nous avons ensuite réalisé des analyses différentielles de méthylation entre un âge de référence (19 mois) et tous les autres (Figure 34B). Des DMC ont pu être identifiées pour chaque comparaison, témoignant de variations fines en fonction de l'âge des animaux. Le nombre de DMC (entre 5 et 29 selon les comparaisons) témoigne d'une faible dynamique de méthylation aux âges analysés (Figure 34C). A l'exception de 4 DMC, les DMC identifiées sont spécifiques d'une comparaison donnée.

En conclusion, on peut remarquer que malgré l'absence de variation globale du méthylome, des variations ont été observées à la résolution du CpG. Néanmoins, ces variations ne concernent que très peu de DMC, suggérant une certaine stabilité du méthylome spermatique aux âges étudiés.

Intersection avec les DMC de la cohorte « fertilité »

Nous avons donc vérifié l'intersection entre les DMC « âge » et les DMC « fertilité » identifiées en race Holstein et décrites plus haut. Nous n'avons identifié que 3 DMC communes. L'âge des animaux, entre 13 et 19 mois, interfère peu avec le niveau de méthylation des DMC liées à la fertilité. Néanmoins, pour éviter tout biais, il sera important de supprimer ces DMC de la liste de biomarqueurs à déployer sur le terrain pour prédire la fertilité en race Holstein.

I.III.III : Analyse de la cohorte « déviation »

Méthylation moyenne entre les deux éjaculats

Enfin, nous nous sommes intéressés à la modification du méthylome spermatique lors de déviations de performances de fertilité pour un même taureau. Nous avons, comme précédemment, analysé le niveau de méthylation moyen entre échantillons déviants et cohérents (Figure 35A). Comme les animaux de la cohorte déviation appartiennent aux 2 groupes de fertilité, nous avons conduit les analyses séparément en fonction de la fertilité des animaux. Que ce soit chez les animaux fertiles ou

subfertiles, les conclusions sont les mêmes, on ne remarque pas de variation du niveau de méthylation moyen du méthylome spermatique entre échantillons déviants et cohérents, ce qui n'exclut pas des variations touchant des CpG individuels.

Analyses différentielles de méthylation

Nous avons dans un second temps réalisé des analyses différentielles par groupe de fertilité entre les échantillons cohérents et déviants (Figure 35B). Le nombre de DMC identifiées est faible (entre 36 et 16), ce qui suggère qu'une baisse rapide de fertilité chez un taureau n'est pas liée à la méthylation de l'ADN. Nous avons également comparé ces dernières et celles identifiées dans l'analyse de la cohorte fertilité en race Holstein. Il n'y a aucune intersection, suggérant une indépendance entre les DMC « fertilité » et les DMC identifiées dans une perte rapide de fertilité.

Conclusion

En conclusion de ce travail, nous avons pu remarquer que l'augmentation de l'âge des taureaux, ainsi que des déviations rapides de la fertilité, étaient corrélées à des variations minimes de méthylation touchant au plus quelques dizaines de CpG. Cela montre que ces deux facteurs ont un impact assez limité sur le méthylome spermatique. Néanmoins, certaines de ces variations pourraient détériorer les modèles construits jusqu'à maintenant. En particulier, nous avons pu constater que 3 DMC étaient communes entre la cohorte âge et la cohorte fertilité. Dans des travaux ultérieurs, il sera important de supprimer ces DMC pour garantir des données non biaisées.

En ce qui concerne les variations qui sont liées à l'âge, les animaux étudiés avaient tous plus de 13 mois, ce qui veut dire qu'ils étaient tous pubères. Il a été montré chez le bovin que des variations du méthylome spermatique étaient plus importantes au moment de la puberté, mais qu'après, le méthylome spermatique se stabilisait (Lambert *et al.*, 2018). Les résultats obtenus dans ce travail sont en adéquation avec ces observations, car les variations observées sont faibles. Il est possible qu'en

étendant cette étude à des stades ultérieurs, plus de DMC puissent être identifiées, de façon comparable aux résultats rapportés par Takeda *et al*.

En ce qui concerne les déviations rapides de fertilité, nous n'avons pas pu les associer à des modifications importantes du méthylome spermatique. Il est connu que la méthylation de l'ADN est une marque plutôt stable comparé aux autres marques épigénétiques. Ainsi, une baisse rapide de fertilité pourrait impliquer d'autre facteurs épigénétiques. On peut également penser que d'autres facteurs que l'épigénétique interviennent dans ce processus.

Ces résultats sont également importants à prendre en considération dans le cadre du développement technologique d'un outil terrain se basant sur la méthylation. Cette partie de discussion est abordée dans la discussion générale de ce manuscrit de thèse.

II : Intégration des données pour une prédiction plus fiable de la fertilité

Dans le chapitre précédent nous avons pu mettre en évidence quelques centaines de DMC en fonction de la race étudiée, entre des éjaculats de taureaux subfertiles et des éjaculats de taureaux fertiles. Basé sur le niveau de méthylation de certaines de ces DMC, il a été possible de construire des modèles de classification avec des performances correctes en race Montbéliarde et très bonnes en race Holstein.

Dans une perspective d'application sur le terrain, il pourrait être intéressant, tout du moins pour la race Montbéliarde, d'intégrer d'autres types de données à la méthylation dans le but d'améliorer la qualité de la prédiction. La plus-value de l'intégration sera moindre en race Holstein du fait des très bonnes performances atteintes avec uniquement la méthylation de l'ADN. Dans les deux cas, l'intégration apporte un éclairage intéressant sur les liens biologiques des différents types de données entre eux et avec la fertilité.

Pour ces raisons, une intégration des données de méthylation de l'ADN spermatique, de sncRNA, de SNP et de paramètres spermatiques a été menée sur les cohortes fertilité des deux races. Les objectifs principaux de ces travaux sont d'analyser les potentielles relations entre les variables en lien avec la fertilité, de construire des modèles de prédiction et d'analyser les variables les plus contributrices de ces modèles afin de proposer des hypothèses biologiques. Comme précédemment, le travail en race Montbéliarde a fait l'objet d'un article qui est actuellement en fin d'écriture et qui sera soumis dans les prochaines semaines. Un court résumé de ce travail est présenté ci-dessous. J'ai ensuite cherché à approfondir les résultats obtenus par une étude très préliminaire d'inférence de réseaux, toujours en race Montbéliarde. Enfin, je présenterai les résultats d'intégration de données obtenus en race Holstein.

II.I : Analyse intégrative en race Montbéliarde

II.I.I : Résumé de l'article 2

Les différentes tables de données de départ (une pour chaque type de données biologiques) sont pour certaines de très grande dimension, avec plus d'un million de CpG pour la méthylation (après filtre des CpG polymorphes), plus de quatre cent mille scnRNA et plus de quarante mille SNP. La seule table de données de faible dimension est celle des paramètres spermatiques avec 11 variables. C'est pourquoi les trois plus grandes tables de données ont été soumises à deux étapes de pré-sélection dans le but de ne conserver que les variables les plus pertinentes. Ainsi, 12 006 variables représentant quatre types de données biologiques en quantité équilibrée (sauf pour les paramètres spermatiques) ont été analysées.

Dans un premier temps, une AFM a montré qu'une partie de la variance de ce jeu de données était liée à la fertilité des animaux. Les variables liées à la fertilité appartiennent aux jeux de données de méthylation de l'ADN, des sncRNA et des SNP. Nous avons ensuite analysé les relations pouvant exister entre ces variables liées à la fertilité. Aucune corrélation significative n'a pu être mise en évidence entre jeux de données, suggérant une structure d'indépendance.

Dans un second temps, des modèles de prédiction ont été construits à partir de quatre approches différentes : les forêts aléatoires, le gradient boosting, la régression logistique avec une pénalité de type Lasso et les réseaux de neurones (Rosenblatt, 1958; Breiman, 2001; Friedman, 2001). L'objectif était de sélectionner les variables permettant d'optimiser la prédiction. Ainsi, à partir d'une centaine de variables pour chaque méthode, le statut de fertilité des animaux a pu être prédit avec une AUC comprise entre 0,9 et 0,99. Cela démontre une bonne capacité de prédiction des modèles, supérieure à celle obtenue avec la méthylation de l'ADN seule.

Enfin les gènes correspondant aux variables les plus contributrices des modèles ont été analysés, mettant en évidence des fonctions décrites dans le développement embryonnaire.

Les figures supplémentaires sont présentées en Annexe 6.

Article 2

Omics data integration for bull fertility prediction

Valentin Costes[1,2,3,4], Eli Sellem[1,2,3], Sylvain Marthey[4,5], Chris Hoze[3,4], Aurélie Bonnet[1,2,3], Laurent Schibler[3], Hélène Kiefer[1,2] and Florence Jaffrezic[4*]

[1]Université Paris-Saclay, UVSQ, INRAE, BREED, 78350 Jouy-en-Josas, France.

[2]Ecole Nationale Vétérinaire d'Alfort, BREED, 94700, Maisons-Alfort, France.

[3]R&D Department, ELIANCE, 149 rue de Bercy, 75012, Paris, France.

[4]Université Paris-Saclay, AgroParisTech, INRAE, GABI, 78350 Jouy-en-Josas, France.

[5]INRAE, MaIAGE, Université Paris-Saclay, 78350 Jouy-en-Josas, France.

*Corresponding author: florence.jaffrezic@inrae.fr

ABSTRACT

Background

Bull fertility is an important economic trait, and the use of subfertile semen for artificial insemination decreases the global efficiency of the breeding sector and leads to environmental impacts. While the analysis of semen functional parameters helps identify severely infertile bulls, no tools are currently available to precisely predict bull fertility and avoid the commercialization of subfertile semen. Since male fertility is a multifactorial phenotype that depends on genetic, epigenetic, physiological and environmental factors, we hypothesized that an integrative analysis of these parameters could help refine the prediction of bull fertility.

Methods

We combined -omics data (genotypes, sperm DNA methylation at CpGs and sperm small non-coding RNAs: sncRNAs) and semen parameters measured on a large cohort of 98 Montbéliarde bulls with contrasting fertility levels. First, a Multiple Factor Analysis (MFA) was

conducted to study the links between the datasets and fertility. Then four methodologies were considered to predict the fertility status and identify potential fertility biomarkers: Logistic Lasso, Random Forests, Gradient Boosting and Neural Networks. Finally, the features selected by these methods were annotated in terms of genes in order to conduct functional enrichment analyses.

Results

The less relevant features in -omics data were first filtered out, and MFA was run on remaining 12,006 features including the 11 semen parameters and a balanced proportion of each type of –omics data. The results showed that unlike the studied semen parameters, the –omics datasets were related to fertility. The different predictive models built from the four biological datasets all showed a high level of prediction accuracy for the bull fertility status, with an AUC (area under the ROC curve) value between 0.9 and 0.99. The selected features as well as their optimal number and the relative contributions of DNA methylation, sncRNAs and single nucleotide polymorphisms (SNPs) were, however, specific to each methodology. In this study, the semen parameters were never identified as predictive variables. The most contributive CpGs, SNPs and miRNAs targeted genes are involved in early development. Interestingly, fragments derived from ribosomal RNAs were overrepresented among the selected features, suggesting their role in male fertility.

Conclusions

These results highlight the benefit of integrating different sources of meaningful biological data to predict complex phenotypes, and provides indications about the performances of four methods when applied to both quantitative and discrete -omics data.

KEYWORDS

Integrative biology, Machine learning, Bull fertility, Sperm epigenome, DNA methylation, small non coding RNA, Single Nucleotide Polymorphism

BACKGROUND

Male fertility is an important economic trait in the bovine industry, and more specifically in the dairy sector where a large share of animals are bred by artificial insemination (AI) using the semen of high genetic merit bulls. The use of subfertile bulls causes economic losses for different actors in the breeding industry, from the breeders that purchase semen, to the breeding companies in charge of its production and commercialization. The use of semen of reduced fertility also delays calving and lactation, requires multiple interventions of artificial insemination technicians, thus decreasing the global efficiency and sustainability of the dairy sector. For these reasons, identifying subfertile bulls and limiting the diffusion of subfertile semen to herds are important challenges for the breeding sector.

During the last decade, many studies have been undertaken to investigate the ability to predict bull fertility based on genotypes or semen functional parameters. The levels of prediction accuracy achieved during these studies were however not sufficient [1,2]. As well as genetics and semen parameters, a wide range of epigenetic mechanisms such as DNA methylation, sncRNAs and the ratio between histones and protamines also contribute to male fertility [2–6]. This is why epigenetic studies have received an increasing attention in the last years in the context of male fertility, with encouraging results [7–14]. However, DNA methylation seems to be insufficient to explain the whole variance related to bull fertility (Costes et al., 2022). Male fertility is a complex and multifactorial phenotype, and considering only one type of biological data at a time may indeed not be enough to achieve good levels of prediction, as many contributing factors are disregarded. To our knowledge, however, studies exploring the potential of several types of epigenetic mechanisms to predict male fertility in combination with genotypes and semen parameters have never been published.

In the last few years, multi-omics integration analyses have gained popularity among researchers, thanks to the development of affordable high-throughput technologies allowing the collection of different types of data from the same biological samples. In order to analyse these data and create interpretable representations, a large number of statistical techniques

have emerged in the field of data integration, which continues to be an active research area [15–18]. Among these methods, exploratory approaches are used to explore the data and perform descriptive analyses. This is the case, in particular, for the Multiple Factor Analysis (MFA) that can help to study the relationships between the different datasets and catch the main sources of variation among them, in order to explore the factors causing this variation [19]. Another important topic addressed in multi-omic data analyses is the identification of biomarkers for the prediction of a specific phenotype. The most popular methods to achieve this goal belong to the Machine Learning tools, such as Random Forests, Gradient Boosting and Neural Networks [20–23].

Despite the abundance of techniques available for data integration, no integrative studies in the field of male fertility have been published so far, probably due to the need for large cohorts and related budget issues, as well as the expertise required to generate and analyse different types of data. In this study, we integrated genetic (SNPs), epigenetic (DNA methylation at CpGs, sncRNAs) and physiological (semen functional parameters) datasets obtained on a unique cohort of 98 bulls, which is quite large in -omics analyses. We report here on the links between these datasets and bull fertility that were explored using MFA. We also built models from these different biological features using four machine learning methods and assessed their performance, and investigated the biological function of the features that were selected. To our knowledge, this study represents the first attempt to integrate different types of biological data in the field of bull fertility.

METHODS

Description of the data

The animal cohort includes 98 French bulls of the Montbéliarde breed that have been classified as fertile or subfertile based on a corrected non return rate (NRR) at 56 days (Supplementary Figure 1). These 98 bulls were commercialized by two different breeding companies: Evajura

(n=43, maintained at Lons-le-Saunier, France) and Umotest (n=55, maintained at Brindas, France), and for all of them 8 to 10 cryopreserved commercial semen straws have been pooled to constitute the biological sample. This biological sample was used to extract genomic DNA and total RNAs from respectively 20 and 40 million sperm cells, and to measure semen functional parameters. DNA methylation was investigated using reduced representation bisulphite sequencing (RRBS) and the analysed CpGs co-localizing with variants listed in the "1000 Bull Genomes" database were filtered out. These steps are precisely described in (Costes et al.; 2022). RNA extraction, scnRNA library preparation and bioinformatics analysis were performed as described [3]. Furthermore, the sncRNA counts were normalized according to library depth using the R package DESeq2 [24]. Genotypes obtained using EuroGMD 50K DNA chip (Illumina) were provided by the breeding companies and represented 40,479 SNPs that successfully passed quality controls. Consequently, each bull was described by 2,003,005 features that belonged to four different biological datasets (40,479 SNPs, 1,548,563 CpGs, 413,952 sncRNAs and 11 semen parameters). A description of each sample, together with the measured semen parameters is provided in Supplementary Table 1.

Data integration by Multiple Factor Analysis

In order to perform an integrated analysis of these data, we first applied an exploratory approach, namely the Multiple Factor Analysis (MFA). This statistical technique consists in establishing a linear combination of features in order to maximise the variance of the table analysed in different principal components, as in the Principal Component Analysis (PCA). The difference is that the MFA technique is able to deal with multi-type data that can be either quantitative or categorical. Furthermore, each value inside a dataset is standardized by the first singular value of this dataset, in order to make the different sets of data comparable whatever their dimensions. The package FactoMineR (v2.4) was used to perform the MFA analyses, and the graphics were obtained with the factoextra (v1.0.7) package [25]. The DNA methylation, sncRNA and semen parameter data were encoded as numerical features and the

SNPs were encoded as categorical features, which precluded their projection on the same variable factor map as the aforementioned numerical features. Fertility, extraction batch and semen collection centers were considered as supplementary categorical features meaning that they did not contribute to the MFA construction. Before the analysis, all numerical features were scaled in order to allow multi-table comparisons.

Feature selection and predictive models

Random Forests, Gradient Boosting, Lasso Logistic regression and neural networks were used to build predictive models and perform feature selection. A brief description of each method is presented below, as well as the parameters used for each analysis.

Logistic Lasso:

The regularised Logistic Lasso was used as the feature of interest is categorical with two modalities, namely “Fertile” / “Subfertile”, and is explained by a large amount of features. The logistic regression is expressed by the equation below to calculate the probability of an individual to be in class 1 depending on the different x_i values.

$$P(1|X) = \frac{e^{b_0+b_1x_1+\cdots+b_jx_j}}{1+e^{b_0+b_1x_1+\cdots+b_jx_j}}$$

Like in linear regression problems, the purpose is to estimate the different coefficients of the model (b_j). This is achieved by finding the coefficients that maximise the likelihood. In this study, as in most -omics analyses, the number of features (>2 millions) was much larger than the number of observations (n=98). We therefore used a Lasso regularisation, and the coefficients of the model were estimated by maximising the likelihood under an L1 penalty [26]. Two hyper parameters have been tuned, which are the lambda (coefficient of regularisation) and the epsilon (tolerance termination criterion).

Random Forest:

Random Forests are an ensemble method based on classification trees that were applied here using the h2o R package (v3.32.1.3) with the function “h2o.randomForest” and the party R package (v1.3.7) with the function “cforest”. This technique constructs K trees for K subsets of individuals and features. Subsets are obtained by bootstrapping the individuals and sampling a number (mtry) of random features of the original dataset for each node calculation. During this study, K was set to 500 and the mtry parameter was calculated as the square root of the number of features in the original dataset. In each subset, trees were constructed recursively from the original node until the last one by splitting the parent node into two child nodes, using the features that discriminate best between fertile and subfertile bulls. The tree stops growing if a node fills one or more of these three conditions: (i) the leaf contains only one individual; (ii) the maximum depth has been reached (which was 20 here); (iii) the split of the node does not improve enough the classification (in a node, the squared error reduction has to be greater than 0.00001). The feature importance in Random Forest was calculated here by looking at the reduction of the squared error before and after a split node. The reduction was attributed to the features that were responsible for the split. These reductions were then summed for each feature in each tree, and it was possible to know which features were the most relevant (the more the value is important, the more the feature is relevant). In this study, both classical and Cforest versions of the Random Forest approach were applied, with respectively the CART and Ctree procedures to build the classification trees [27,28]. Briefly, the CART procedure selects the features that minimise the Gini impurity criterion, while the Ctree procedure first identifies an explicative feature that is correlated with the variable of interest before choosing the best split inside this feature that minimises the Gini impurity criterion.

Gradient Boosting:

Gradient boosting is also an ensemble method based on classification trees, but the strategy used to build the trees is different from the one used by Random Forests. Indeed, in contrast with Random Forests where each tree is constructed independently from the others, tree (n+1)

is built from the “error” (the residual) of tree (n) in gradient boosting. More details about the precise algorithm of the gradient boosting can be found in Hastie et al. [29]. There were three parameters to optimise in this method, which were the learning rate, the maximum depth of a tree and the number of trees. This method was implemented using the h2o package (v3.32.1.3), with the function h2o.gbm.

Neural Network:

The Neural Network method is based on a combination of artificial neurons that are distributed in different layers. An artificial neuron is described by the mathematical formula below where x_i is the output of neuron i, w_i is the weight given to the output of neuron i, φ is the activation function. The w_o weight is associated with a fictive feature x_o that is equal to 1.

$$\varphi\left(\sum_{i=0}^{m} w_i x_i\right)$$

Each layer can contain a different number of neurons, connected so that each neuron of layer (n+1) is connected to all neurons of layers n and (n+2). The deep learning network is calibrated to do the prediction by estimating the different weights (w_i) by a process called backpropagation of the gradient which is done after each sample passes through the neural network [30]. The Neural Network algorithm was applied here using the h2o package (v3.32.1.3), with the function h2o.deeplearning. The relative feature importance calculation was directly implemented in the package and used the Gedeon method [31].

These four different methods were applied to a matrix that contained the values of DNA methylation at CpGs, sncRNAs expression, genotypes at SNPs and semen parameters for each individual. The animal cohort was split into a training and a testing set. The training set contained 2/3 of the samples (n=65) randomly selected but with a conservation of the original proportion of fertile and subfertile samples, and was used to create the model. The testing set contained the remaining samples (n=33, so 1/3 of the cohort) to evaluate the predictive ability

of the model previously constructed on the training set. This process was iterated 50 times with a resampling of the training and the testing sets and the AUC was averaged over the 50 AUC values obtained for each iteration [32].

Feature annotation and enrichment analysis

Gene annotation of CpGs and SNPs was performed relative to gene features with an in-house pipeline as described (Costes et al., 2022). The reference files were downloaded from Ensembl (ftp://ftp.ensembl.org/pub; release 95). The following criteria were applied: TSS, -100 to +100 bp relative to the transcription start site (TSS); promoter, -2000 to -100 relative to the TSS; TTS, -100 to +100 relative to the transcription termination site (TTS). Genes containing CpGs and/or SNPs in intragenic regions, upstream (up to -10 kb from the TTS) or downstream (up to +10 kb from the TTS) regions were subjected to an enrichment analysis using DAVID (version 6.8) with default parameters and using all genes (n=20,641) targeted by the analysed SNPs and CpGs as a reference.

The gene targets of the miRNAs were identified using TargetScan (version 7.2) with default parameters and the enrichment analysis was performed on the identified genes using WebGestalt against the whole genome.

RESULTS

Data preparation and extraction of relevant features

In order to achieve data integration and construct a model predictive of bull fertility status, we considered four biological datasets relevant to male fertility: sperm DNA methylation and sncRNAs, semen parameters (SPs), and genotypes of the bulls. These different tables included 1,548,563 CpGs (DNA methylation features), 413,952 sncRNAs, 11 SPs and 40,479 SNPs (genotype features), respectively (Figure 1A).

The first step of the analysis was to pre-filter the DNA methylation, sncRNA and SNP data tables in order to remove unusable or unnecessary features and limit the dimension of these tables (Figure 1B). The DNA methylation table included a large number of missing values resulting from uncovered CpGs, or CpGs covered by less than 10 reads and from which DNA methylation values could not be accurately estimated. As some of the methods we used cannot handle missing values, every CpG that contained at least one missing value across the 98 samples was filtered out, which left 641,306 CpGs with no missing value. Assuming that CpGs with extreme methylation values showed little inter-individual variability, we then removed the CpGs that were consistently hypo- (0% to 20%) or hypermethylated (80% to 100%) in all the samples; and from the 98,203 remaining CpGs, we selected the 40,000 CpGs with the highest variance. The sncRNA table also contained a very large number of features, many of them being expressed below detection level in most samples. Because these expression traces could not be thoroughly quantified in all samples, the corresponding sncRNAs were disregarded in further analyses. Only sncRNAs that displayed an average read count above 10 after normalization were therefore kept, which represented 24,172 features in total. Finally, SNPs at which the genotype was identical for all 98 bulls were filtered out, resulting in 38,853 SNPs with polymorphism in our cohort.

In contrast with genotypes that are routinely obtained using standardized procedures, DNA methylation, sncRNA and SP data were acquired in the lab from the 98 semen samples processed in different batches. We already demonstrated that DNA methylation was not significantly affected by the batch (Costes et al., 2022), and this was also confirmed in the current study (Supplementary Figure 2). However, a PCA run on the 24,172 selected sncRNAs and 11 SPs revealed that the batch had a huge impact on these data (Supplementary Figure 3A-B); and that this effect was confounded with the origin of the bulls in terms of semen collection centres. We therefore corrected the data for the batch effect. This correction was carried out using a generalized linear model for the sncRNAs and a linear model for SPs, with the experimental batch as a fixed effect. The residuals of the model were then extracted. The

corrected data were no longer biased according to the batch nor the centre (Supplementary Figure 3C-D, Figure 1C), and could therefore confidently be used for further analyses. This correction for the batch effect did not apply to the SNP data, which are routinely obtained using standardized procedures. Centres 1 and 2 were however distinguished on first dimension of the PCA (Supplementary Figure 4), which reflects that these two centres probably select and commercialize slightly different genetics.

As shown in Supplementary Figures 2, 3C-D and 4, the largest part of the variance in the 24,172 sncRNA, 11 SP, 40,000 CpG and 38,853 SNP data was unrelated to fertility. Therefore, the last step of data preparation consisted in the selection of features relevant to male fertility in a supervised way (i.e., select features that could differentiate fertile from subfertile bulls), using a Random Forest approach (Figure 1D). Due to the small number of features in the SP dataset, only the DNA methylation, sncRNA and SNP data were subjected to this step, which resulted in the selection of 3,851 CpGs, 3,504 SNPs and 4,640 sncRNAs. The selection was based on the relative importance of each feature in the model (Supplementary Figure 5).

At the end of these pre-processing steps, a table containing 12,006 features originating from four different biological datasets, with quite balanced proportions (except for SPs) was therefore obtained and used in the following analyses. These features are provided in Supplementary Table 2-4.

Integration by Multiple Factor Analysis

Next we investigated the potential links between the different types of data and their relationship with male fertility. To this aim, we analysed the four datasets displayed in Figure 1D using MFA.

The global factor map (Figure 2A) shows which types of features contributed to dimensions 1 and 2 among SPs, sncRNAs, CpGs and SNPs, and indicates the positioning of supplementary variables (fertility, semen processing batch, origin of the bulls in terms of semen collection

centres). The CpG, sncRNA and SNP datasets all exhibited high coordinates on the first dimension, meaning that they represent the most important sources of variation and are the main contributors to the first component of the MFA. The SPs, on the other hand, did not contribute much to the first dimension but heavily to the second one. Fertility exhibited a high coordinate on the first dimension and a very low one on the second dimension, meaning that the most important sources of variation in the CpG, sncRNA and SNP datasets are correlated to the fertility status of the bulls. This result was expected given that MFA was conducted on features preselected based on their ability to discriminate fertile and subfertile bulls. On the other hand, fertility was not associated with the second dimension, suggesting that in the studied cohort the fertility status of the animals is independent from the SPs. Finally, and consistent with the correction of the data for these effects, the bulls' origin and the batch did not contribute significantly to the two first dimensions of MFA. As expected given the correlation of fertility with the first dimension, fertile and subfertile bulls were discriminated along this dimension (Figure 2B).

Correlation structure among the CpG and sncRNA features

Because quantitative and qualitative features cannot be displayed on the same variable factor map of the MFA, SNPs were not considered for the correlation analysis. In agreement with the above observations, features belonging to the CpG and sncRNA datasets contributed to the first dimension of MFA while SPs contributed to the second one (Figure 2C). Both features with positive and negative coordinates on the first dimension of the factor map were found, indicating positive and negative correlations with the fertility status. To identify the most relevant features, two clusters were defined based on the most extreme coordinates on first dimension. Cluster 1 contained both CpGs and sncRNAs (170 CpGs and 83 sncRNAs) that were more methylated or expressed in subfertile bulls than in fertile bulls (subfertile bulls having positive coordinates on the individual factor map; Figure 2B) and cluster 2 mostly contained sncRNAs (120 sncRNAs and only three CpGs) that were more expressed in fertile

bulls than in subfertile bulls. Interestingly, most of the scnRNAs belonging to these clusters were sncRNAs of the miRNA and rRF (ribosomal RNAs derived fragment) families (Supplementary Figure 6).

The potential correlation structure among features belonging to cluster 1 was then investigated as CpGs and sncRNAs were represented in equivalent proportions. By looking at the feature graph on Figure 2C it appeared that arrowheads from CpGs and scnRNAs were close to each other's, suggesting a correlation structure among these two types of biological features. However, the features were not highly correlated with the first dimension (cos^2 with the first dimension was not high), and the arrowheads could be far from each other on other dimensions, as illustrated using the first and third dimensions where the orthogonal structure between sncRNAs and CpGs suggest very limited correlations (Figure 2D). This was further confirmed by the correlation heatmap (Figure 3) showing the absence of correlation structure between the CpGs and scnRNAs, while positive correlations were observed within each dataset of cluster 1.

In summary to this part, results of MFA demonstrated that genotypes, sperm sncRNA expression and DNA methylation all individually contributed to fertility, with limited correlations between features of different biological types.

Optimisation of the hyper-parameters of the predictive models

The next step was to build predictive models for the bull fertility status using the most relevant biological features. For that purpose, we considered four different methodologies: Random Forest, Logistic Lasso, Gradient Boosting and Neural Networks (Figure 4A), each using hyper-parameters that had to be calibrated (Figure 4B-C).

In the case of Random Forest, the recommended parameters (500 trees; mtry = square root of the number of features present in the model) were used since they proved to be robust [33] (Figure 4B). For the three other methods, however, there are no consensual parameters and

they have to be tuned on a case-by-case basis. To this aim, a grid of different values was set for every parameter. All parameter combinations were evaluated, which represented 42, 90 and 72 combinations for Logistic Lasso, Gradient Bosting and Neural Networks, respectively, and those that maximised the AUC were chosen for further analyses (Figure 4C). The Logistic Lasso parameters were assessed using a 10 fold cross validation, which is recognized as the most robust approach for parameter calibration. This approach could not be applied to Gradient Boosting and Neural Networks, which both included more parameters to assess and required extensive calculation time. For these methods, the calibration of parameters was therefore performed using one training and one testing set. For Logistic Lasso a grid of [0.01, 0.1, 1, 5, 7, 10] was considered for the lambda parameter (regularisation parameter) and [0.0005, 0.005, 0.001, 0.05, 0.1, 0.15, 0.2] for epsilon (tolerance parameter), which corresponded to 42 different combinations (Figure 4C). For Gradient Boosting a grid of [1, 3, 5, 7, 9, 11] was considered for the maximum depth of the tree, [500, 1000, 1500] for the number of trees and [0.1, 0.2, 0.3, 0.4, 0.5] for the learning rate parameter, which corresponded to 90 different combinations. Finally, for Neural Networks the hidden layers architectures considered were [(200, 100, 50), (100, 50, 25), (100, 50)], which corresponded to two or three layers, and a number of neurons per layer varying from 25 to 200. The grid for the learning rate was [0.05, 0.1, 0.2], and [0, 0.1,1, 5] for the lambda value of the L1 penalty in the cost function. Two activation functions were evaluated, namely Rectifier and Tanh, resulting in 72 combinations. In total for the three methods, all different combinations were tested and the chosen combinations were those maximizing the AUC values (Figure 4D).

Comparison of several methods for feature selection and fertility prediction

After calibration of the different parameters, the prediction models were applied to the four datasets, with two goals: compare the prediction ability of the different methods regarding the bull fertility status and identify the most relevant features.

For each method, a model was first constructed using the 12,006 features. The number of features included in the model was then progressively decreased, and the AUC value was calculated in each case (Figure 5, left panel). Interestingly, each method exhibited a specific behaviour regarding the optimal number of features. For example, Neural Networks showed poor performances with small numbers of features (AUC = 0.74 with 10 features), and greatly improved when including all the features in the model (AUC = 0.9). To the opposite, Gradient Boosting outperformed with a small number of features in the model (AUC=0.88 with 10 features) but was not robust to a large number of features (AUC=0.65 with all features). Most importantly, and despite these different behaviours, all methods were able to accurately predict bull fertility with specific numbers of features, as illustrated by the minimal number of features that maximised the AUC for each method (Figure 5, right panel). The Random Forest method showed the lowest performance with an AUC of 0.9 (which is still quite good) and 100 features. The highest performance was reached by the Neural Network with an AUC of 0.99, which is excellent, but to the expend of a large number of features (500 features, compared to only 50 for Gradient Boosting with an AUC of 0.92).

In conclusion, these results demonstrated that a thorough calibration of the parameters associated with each method, followed by an optimization of the number of features included in the model, allowed the identification of subfertile bulls whatever the method used. Although differences were observed in terms of optimal numbers of features, the comparable performances obtained with the four methods also suggest that the 12,006 features contain sufficient information to robustly predict bull fertility.

Nature of the selected features

To get insight into the information critical for fertility prediction that was embedded in the four datasets, we studied more in detail the features selected by the four methods, with a particular attention to their biological nature.

To this aim, we first compared the optimal features selected by the four methods (Figure 6A). Strikingly, only two features were common to all four methods, demonstrating the specificity of each method. A certain degree of overlap could nevertheless be observed. Indeed 66% of the features selected by the Random Forest approach were also selected by at least one other method, and this percentage reached 88% for Gradient Boosting. In contrast, the methods using larger numbers of features did not share many features with other models (26% for Logistic Lasso and only 3% for Neural Network).

The nature of the selected features was then analysed for each method (Figure 6B). Whatever the method, the SPs were never identified as important features for fertility prediction. This result is in agreement with the absence of correlation between SPs and bull fertility observed in the MFA. The Logistic Lasso, Random Forest and Gradient Boosting behaved quite similarly, as they selected mainly CpGs and sncRNAs (Figure 6B, left panel). It has to be noted that the CART procedure used for the construction of classification trees in the Random Forest and Gradient Boosting methods is known to be biased toward the selection of continuous features rather than qualitative features [28]. Given the distribution of the –omics features (continuous for CpGs and sncRNAs vs. discrete for SNPs), this property of the CART procedure might explain why very few SNPs were selected. However, it should be mentioned that this selection bias has never been reported for the Logistic Lasso which also favoured the selection of CpGs and sncRNAs over SNPs. Interestingly, the cforest method, which is related to Random Forest but uses the Ctree procedure instead of CART to build the classification trees, was shown to be unbiased toward the data type and distribution [28]. This method therefore offered an opportunity to assess the effect of the CART procedure on the nature of the selected features, at least for Random Forest, and was applied to the 12,006 features as described above. The number of features that maximised the model performance was 100, similar to the standard Random Forest approach, but the prediction accuracy was slightly better (AUC of 0.92 compared to 0.90; Supplementary Figure 6). The pattern of the selected features was different from the classical Random Forest approach, with a much larger

proportion of selected SNPs (Figure 5B, upper right panel), confirming that the CART procedure biased feature selection. Finally, the Neural Network approach displayed another pattern of selection, where the SNPs represented the main type of selected features (Figure 5B, lower right panel).

In conclusion, although each method behaved specifically in terms of number and nature of the selected features, they all selected at least two types of features, highlighting the benefit of data integration for the prediction of bull fertility. In addition, all types of –omics features were selected, suggesting that they were all relevant to bull fertility prediction.

Functional annotation of the selected features

In order to investigate if the features contributing to fertility predictions are biologically relevant, we annotated them regarding genes and sncRNA families, with a focus on the features selected by unbiased methods (Logistic Lasso, cforest and Neural Networks)..

We used different strategies depending on the nature of the selected features (Figure 7A). The CpG and SNP features were analysed together since they both have the potential to target coding genes directly, affecting the regulation of their expression. Thus, the genes containing selected CpGs or SNPs in their body or flanking regions were first identified and then subjected to functional enrichment analysis. The situation was more complex for sncRNAs that have various functions according to the family they belong to [3]. Because tools allowing the *in silico* identification of putative miRNA targets are available, we sought the genes potentially targeted by the selected miRNA features and then subjected them to functional enrichment analysis [34].

When each method was analysed separately, no individual gene ontology (GO) term was found significantly enriched among genes containing selected CpG and SNP features in any method. One and two clusters with significant enrichment scores (EASE scores above 1.3) were however found using DAVID, from the CpG and SNP features selected by Neural

Networks (164 gene IDs, Supplementary Figure 8A) and Logistic Lasso (64 gene IDs, Supplementary Figure 8B), respectively. In contrast, no enriched cluster could be found from the features selected by cforest that covered only 31 gene IDs. The enriched clusters found with Neural Networks and Logistic Lasso displayed no overlap in either genes or terms. Similarly, the distributions of sncRNA families among the selected sncRNA features greatly differed between the three methods (Supplementary Figure 9), demonstrating the specific behaviour of each method regarding the types and functions of the selected features.

We next grouped the information of the three unbiased method in order to increase the number of targeted genes and thus the relevancy of the functional enrichment analysis. This led to a list of 787 unique features (Supplementary Table 3). CpG and SNP features could be associated to 319 gene IDs that did not show any significant enrichment for individual GO terms. Three clusters with a significant enrichment score were however found using DAVID, that respectively gathered homeobox-containing genes, genes involved in ionic transport and proteins containing a Fibronectin domain (Figure 7B). The homeobox cluster is of particular interest because the genes of this family play key roles in embryonic and foetal development [35]. The fibronectin cluster is also compelling since fibronectin is a glycoprotein important to spermatozoa physiology and has a role in the interaction between gametes [36]. Noteworthy, CpG features mainly targeted the homeobox and ionic transport clusters, whereas the SNPs were mainly annotated to the fibronectin cluster.

We also compared the distributions of the different sncRNA families among the selected features and the background. The rRFs and miRNAs were enriched among the selected features, while PIWI-associated RNAs (piRNAs) were depleted and tRFs (transfer RNAs derived fragments) remained unchanged (Figure 7A). A total of twelve miRNA features were selected by the three methods. Their putative target genes (5601 gene IDs) were identified *in silico* using TargetScan, based on homologies with the miRNA seeds [34]. An enrichment analysis of these putative targets was then conducted using WebGestalt (Figure 7C).This method was used, because DAVID cannot process gene lists containing more than 3000

genes. GO terms related to cell differentiation, embryonic and foetal development were overrepresented, such as: “Regulation of cell morphogenesis”, “Morphogenesis of an epithelium”, “Synapse organization”, “Axon development”, “Regulation of neuron projection development”, “Morphogenesis of a branching structure” and “Organ growth”. These results therefore suggest that the selected miRNA features potentially regulate genes involved in important developmental processes.

In conclusion, the most predictive features selected by the three methods were annotated to genes and sncRNA families potentially important to development. Alteration of their status in subfertile bulls may therefore have subtle effects on gene expression after fertilization, resulting in altered developmental outcomes with indirect negative impact on male fertility.

DISCUSSION

The purpose of this study was to analyse the links between semen parameters, DNA methylation, sncRNA expression, genetic polymorphism and male fertility of 98 AI bulls with contrasted fertility levels using data integration methodologies. MFA was run on a total of 12,006 features obtained after data filtering, processing, and a first round of feature selection, which highlighted the CpGs, sncRNAs and SNPs as the main sources of variability correlated to the fertility status. Different methodologies were then used to construct predictive models with quite good performances from these pre-selected features. Finally, the features that contributed the most to these models were represented by at least two types of –omics data and were related to genes relevant to male fertility and development.

DNA methylation and sncRNA expression independently contribute to bull fertility

The first interesting finding was that bull fertility was not linked to the semen parameters analysed during this study, which were related to the viability, motility and mitochondrial status of spermatozoa [13]. As previously shown in our previous research each semen sample a large share of sperm was viable post thawing, with a correct motility and enough energy provided

by the mitochondria to sustain its function. The absence of major alteration of semen parameters related to subfertility underscores the importance of the research efforts conducted to identify potential causes of idiopathic subfertility and infertility with a normozoospermic profile in both livestock species and humans [5,37]. In contrast, features among the CpGs, sncRNAs and SNPs were all linked to fertility, which was further confirmed by the modelling and feature selection approaches. Correlations between the quantitative features (CpGs and sncRNAs) that were best represented on the first dimension of MFA did exist within a single dataset but not between CpG and sncRNA datasets, an observation also confirmed using partial least square analysis (data not shown). Although correlations between these two types of epigenetic processes could be somehow expected [38], the result we obtained is not consistent with the existence of direct regulations of sncRNA expression by DNA methylation. However, the peculiar transcriptional status of sperm cells should be taken into account when interpreting this result. Indeed, sperm cells are transcriptionally inactive, and most of the sncRNAs accumulated during spermatogenesis are actually piRNAs [3,39]. The sperm RNA content is then drastically modified during the transit of sperm through the epididymis, where rRFs, miRNAs and tRFs are gained. These sncRNAs gained during the post-testicular maturation are not transcribed by the spermatozoa but transferred through epididymosomes. Their expression is therefore independent from the sperm DNA methylome. Furthermore, some studies suggest that sncRNAs gained through epididymis are important to the first stages of embryogenesis, while on the other hand, piRNAs are important for fertility mainly during the spermatogenesis stages [40,41]. Here, we selected scnRNAs that are linked to fertility, but as we did not observe major alterations of semen parameters, it could be speculated that part of the molecular events related to subfertility arose after spermatogenesis, involving sncRNAs acquired during the epididymal transit rather than piRNAs transcribed during earlier stages. This hypothesis is in line with the observation that most of the scnRNAs identified as related to fertility during the MFA belong to the miRNA and rRF families.

Predictive performances of the models and selected features

An interesting aspect highlighted by our study is the precision of bull fertility prediction using a combination of –omics data. Indeed, whatever the method applied, the AUC varied between 0.9 to 0.99 using a number of features comprised between 50 and 500, demonstrating the accuracy and sparsity of all models investigated. These results contrast with our previous findings (Costes et al., 2022], where bull fertility was predicted with an AUC of 0.8 on the same cohort, using only DNA methylation and a Random Forest approach. A closer look at the 25-30% misclassified bulls then demonstrated that they displayed a DNA methylation profile typical of the opposite fertility class at selected CpG features, explaining why the predictive accuracy was lower than during the current integrative analysis. Here, most of these bulls were allocated to the correct fertility class, showing that additional -omics information was necessary to improve the prediction accuracy. As the selected features includes a mixture of SNPs, sncRNAs and CpGs for each model, we conclude that in agreement with the multi-factorial nature of male fertility, several types of –omics data may be necessary to predict bull fertility.

Another interesting aspect is that, although all models exhibited good performances, only few features were selected in common, and the nature of the selected features also greatly differed as a function of the method applied. Importantly, Logistic Lasso is an additive method, while the other methods do not make any assumption about the relationships between features and can deal with complex non-linear interaction patterns. Furthermore, the importance of features in Logistic Lasso is assessed individually for each feature, a property shared with genome wide associations studies (GWAS) that usually investigate the genetic association between individual SNPs and a given phenotype. Of note, only few SNPs have been so far identified by GWAS as belonging to male fertility quantitative trait loci (QTLs) [42]. This could be explained by the fact that male fertility is a complex trait with a relatively low heritability. Rather than a major gene, several SNPs in combination, each explaining a small part of the phenotypic variance, could then be involved in male subfertility (at least for normospermic cases), limiting the identification of QTLs by classical GWAS methods. Similarly, because of its reduced ability

to select relevant combinations of SNPs, Logistic Lasso may preferentially select sncRNAs and CpGs that could potentially have a larger individual influence on fertility than SNPs. This behaviour could therefore possibly explain the huge proportion of selected sncRNAs and CpGs (242 out of 250) with respect to SNPs. The reason why Gradient Boosting and classical Random Forest also selected more sncRNAs and CpGs than SNPs is probably different. Indeed, unlike Logistic Lasso these two methods are not purely additive and can select features with possible interactions, but they are both biased toward the selection of quantitative rather than qualitative features [28]. We therefore switched to the cforest method, which is a Random Forest approach able to account for this bias, leading to the selection of each type of -omics data in quite balanced proportions. Finally, the Neural Network approach, which is described for dealing with complex interactions among features, selected the largest proportion of SNPs. Based on this result, it can be hypothesized that while CpGs and scnRNAs individually contribute to fertility, a defined combination of SNPs displaying a complex correlation structure that is best accounted by Neural Networks, is necessary to reach the maximal prediction accuracy in this bull cohort. This hypothesis is aligned with studies where machine learning methods outperformed classical methods of genetic evaluation (GBLUP or Bayesian Model), especially when the relationships among SNPs are not purely additive [43].

Which method is the best for fertility prediction?

This study was conducted in an integrative context, mixing together data from different origins and distributions, which affected the performances of the five investigated methods at different degrees. As mentioned above, classical Random Forest and Gradient Boosting were both biased regarding the distribution and nature of the data, which did not directly influence the model performance but clearly affected feature selection. Among the three remaining methods, Logistic Lasso is not described for displaying such a selection bias, but it evaluates feature contribution one by one without taking into account non-additive interactions. This property could possibly be a drawback because in -omics integrative studies the different types of data can possibly interact in a non-additive way. The two remaining methods, cforest and Neural

Network, rely on different mathematical principles but share two properties: (i) they are not sensitive to data origin or distribution, and (ii) they both allow complex interactions between features without any *a priori.* They are therefore both well suited for prediction and feature selection in integrative analyses of –omics data. However, cforest is easier to handle than Neural Networks, with very few hyper parameters to optimize and also relevant recommended parameters. Oppositely, even though an artificial neuron is a simple mathematical object, Neural Networks are very complex with a lot of hyper parameters to optimize and many possible architectures (autoencoder, convolutional etc). Their successful application requires both background knowledge and computing resources. This is precisely the reason why cforest is probably best suited for biologists that wish to build predictive models and perform feature selection from –omics data.

Epigenetic features and genetic features target different functions

Strikingly, genes involved in embryonic development were overrepresented among the putative target genes of selected miRNAs. Although it can be argued that the use of a software for the *in silico* identification of miRNA target genes can lead to false positives, it is noteworthy that some miRNAs identified during this study, such as miR-100 and miR-29a that are both ranked at the top of the selected miRNA features, have well established functions in embryonic development [44,45]. In addition, miR-339a, miR-449a, mir-1246 and miR-21-5p, all identified during this study, have also been highlighted as differentially expressed between high and low fertility bulls [46,47], suggesting their relevancy to bull fertility. Finally, the selected rRF features included 18S, 12S, 28S, 16S and 5.8S subspecies. The rRFs have long been regarded as degradation products of rRNAs without any biological significance. As rRFs are acquired by sperm during the transit through epididymis, they could potentially be important to embryonic development. Moreover, they have been found to interact with the AGO protein, suggesting a role in gene regulation like miRNAs [48]. Due to a lack of knowledge about the role of rRFs in fertility, this hypothesis remains speculative. It can be pointed out that one study reported a

differential expression for some rRFs between two groups of patients with contrasted IVF outcomes [49].

A cluster of terms related to homeobox genes was significantly enriched among the selected CpG features. This cluster included *PBX1, BARHL1, MKX, LHX3, ALX4, ZFHX4, HOXB1, TLX3* and *HMX1* homeobox genes that are involved in different steps of the embryonic and foetal development [50–58]. This cluster also contained two genes with a CXXC-type zinc finger motif (*CXXC1* and *CXXC5)*, and three transcription factors (*NFIX*, *TEAD2* and *ELF2)* that all play an important role in the embryonic development [59–63]. Another cluster of interest is related to fibronectin, which is a glycoprotein involved in cell adhesion to the extracellular matrix. It is located at the head of spermatozoa and facilitates the interaction between spermatozoa and the oocyte [36]. The proteins included in this cluster have so far not been reported to mediate gamete interactions, and contain a fibronectin type III domain, which is involved in cell-to-cell interactions.

Interestingly, the epigenetic mechanisms target embryonic development genes. Subtle changes in the expression of developmentally important genes by an altered DNA methylation profile of the paternal genome and/or an altered sperm sncRNA content may potentially affect the normal course of development and result in increased losses of embryos, finally influencing male fertility without dramatic consequences. In contrast, the genetic transmission of developmentally unfavourable alleles may have more severe impacts on the gestation outcomes and lead to more extreme phenotypes that are not represented in the present cohort.

CONCLUSIONS

Based on four kinds of methodologies (Logistic Lasso, Random Forests/cforest, Gradient Boosting and Neural Networks), the bull fertility status could be predicted from a subset of explicative features with high precision, demonstrating the benefit of integrating different –omics datasets to predict complex phenotypes such as male fertility. The subset of selected

features was unique to each method, which could be related to the specific behaviours of the methods towards feature distribution and correlation structure. A common characteristics shared by all the methods is that the epigenetic features that were selected (both CpG DNA methylation and sncRNA expression) pointed to embryonic development as the main process potentially dysregulated in our cohort. This result calls for the generation of embryos with the semen samples used during this study in order to dissect the underlying molecular mechanisms.

REFERENCES

1. Sellem E, Broekhuijse MLWJ, Chevrier L, Camugli S, Schmitt E, Schibler L, et al. Use of combinations of in vitro quality assessments to predict fertility of bovine semen. Theriogenology. 2015;84:1447-1454.e5.

2. Taylor JF, Schnabel RD, Sutovsky P. Review: Genomics of bull fertility. Animal. 2018;12:s172–83.

3. Sellem E, Jammes H, Schibler L, Sellem E, Jammes H, Schibler L. Sperm-borne sncRNAs: potential biomarkers for semen fertility? Reprod Fertil Dev

4. Boissonnas CC, Jouannet P, Jammes H. Epigenetic disorders and male subfertility. Fertility and Sterility. 2013;99:624–31.

5. Carrell DT. Epigenetics of the male gamete. Fertil Steril. 2012;97:267–74.

6. Cho C, Willis WD, Goulding EH, Jung-Ha H, Choi YC, Hecht NB, et al. Haploinsufficiency of protamine-1 or -2 causes infertility in mice. Nat Genet. 2001;28:82–6.

7. Kropp J, Carrillo JA, Namous H, Daniels A, Salih SM, Song J, et al. Male fertility status is associated with DNA methylation signatures in sperm and transcriptomic profiles of bovine preimplantation embryos. BMC Genomics. 2017;18:280.

8. Verma A, Rajput S, De S, Kumar R, Chakravarty AK, Datta TK. Genome-wide profiling of sperm DNA methylation in relation to buffalo (Bubalus bubalis) bull fertility. Theriogenology. 2014;82:750-759.e1.

9. Takeda K, Kobayashi E, Ogata K, Imai A, Sato S, Adachi H, et al. Differentially methylated CpG sites related to fertility in Japanese Black bull spermatozoa: epigenetic biomarker candidates to predict sire conception rate. J Reprod Dev. 2021;

10. Gross N, Peñagaricano F, Khatib H. Integration of whole-genome DNA methylation data with RNA sequencing data to identify markers for bull fertility. Anim Genet. 2020;51:502–10.

11. Fang L, Zhou Y, Liu S, Jiang J, Bickhart DM, Null DJ, et al. Comparative analyses of sperm DNA methylomes among human, mouse and cattle provide insights into epigenomic evolution and complex traits. Epigenetics. 2019;14:260–76.

12. Narud B, Khezri A, Zeremichael TT, Stenseth E-B, Heringstad B, Johannisson A, et al. Sperm chromatin integrity and DNA methylation in Norwegian Red bulls of contrasting fertility. Mol Reprod Dev. 2021;88:187–200.

13. Costes V, Chaulot-Talmon A, Sellem E, Perrier J-P, Aubert-Frambourg A, Jouneau L, et al. Predicting male fertility from the sperm methylome: application to 120 bulls with hundreds of artificial insemination records. Clin Epigenetics. 2022;14:54.

14. Štiavnická M, Chaulot-Talmon A, Perrier J-P, Hošek P, Kenny DA, Lonergan P, et al. Sperm DNA methylation patterns at discrete CpGs and genes involved in embryonic development are related to bull fertility. BMC Genomics. 2022;23:379.

15. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics Data Integration, Interpretation, and Its Application. Bioinform Biol Insights. 2020;14:1177932219899051.

16. Eicher T, Kinnebrew G, Patt A, Spencer K, Ying K, Ma Q, et al. Metabolomics and Multi-Omics Integration: A Survey of Computational Methods and Resources. Metabolites. 2020;10.

17. Tini G, Marchetti L, Priami C, Scott-Boyer M-P. Multi-omics integration—a comparison of unsupervised clustering methodologies. Briefings in Bioinformatics. 2019;20:1269–79.

18. Picard M, Scott-Boyer M-P, Bodein A, Perin O, Droit A. Integration strategies of multi-omics data for machine learning analysis. Comp Struct Biotechnol J. Amsterdam: Elsevier; 2021;19:3735–46.

19. Becue-Bertaut M, Pages J. Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data. Comput Stat Data Anal. Amsterdam: Elsevier Science Bv; 2008;52:3255–68.

20. Breiman L. Random Forests. Machine Learning. 2001;45:5–32.

21. Mason L, Baxter J, Bartlett P, Frean M. Boosting Algorithms as Gradient Descent. Advances in Neural Information Processing Systems [Internet]. MIT Press; 2000 [cited 2021 Nov 4]. Available from: https://proceedings.neurips.cc/paper/1999/hash/96a93ba89a5b5c6c226e49b88973f46e-Abstract.html

22. Friedman JH. Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics. Institute of Mathematical Statistics; 2001;29:1189–232.

23. Rosenblatt F. The Perceptron - a Probabilistic Model for Information-Storage and Organization in the Brain. Psychol Rev. Washington: Amer Psychological Assoc; 1958;65:386–408.

24. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.

25. Le S, Josse J, Husson F. FactoMineR: An R package for multivariate analysis. J Stat Softw. Los Angeles: Journal Statistical Software; 2008;25:1–18.

26. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996;58:267–88.

27. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification And Regression Trees. Boca Raton: Routledge; 2017.

28. Strobl C, Boulesteix A-L, Zeileis A, Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics. London: Bmc; 2007;8:25.

29. Hastie T. Tibshirani R. Friedman J.: The Elements of Statistical Learning. Springer-Verlang (es); 2001.

30. Werbos PJ. Backpropagation through time: what it does and how to do it. Proceedings of the IEEE. 1990;78:1550–60.

31. Gedeon TD. Data Mining of Inputs: Analysing Magnitude and Functional Measures. Int J Neur Syst. 1997;08:209–18.

32. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006;27:861–74.

33. Probst P, Boulesteix A-L. To tune or not to tune the number of trees in random forest. J Mach Learn Res. 2017;18:6673–90.

34. Agarwal V, Bell GW, Nam J-W, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. Izaurralde E, editor. eLife. eLife Sciences Publications, Ltd; 2015;4:e05005.

35. Mark M, Rijli FM, Chambon P. Homeobox genes in embryogenesis and pathogenesis. Pediatr Res. 1997;42:421–9.

36. Hoshi K, Sasaki H, Yanagida K, Sato A, Tsuiki A. Localization of fibronectin on the surface of human spermatozoa and relation to the sperm-egg interaction. Fertil Steril. 1994;61:542–7.

37. Åsenius F, Danson AF, Marzi SJ. DNA methylation in human sperm: a systematic review. Hum Reprod Update. 2020;26:841–73.

38. Moore LD, Le T, Fan G. DNA Methylation and Its Basic Function. Neuropsychopharmacology. 2013;38:23–38.

39. Sellem E, Marthey S, Rau A, Jouneau L, Bonnet A, Perrier J-P, et al. A comprehensive overview of bull sperm-borne small non-coding RNAs and their diversity across breeds. Epigenetics Chromatin. 2020;13:19.

40. Weick E-M, Miska EA. piRNAs: from biogenesis to function. Development. 2014;141:3458–71.

41. Conine CC, Sun F, Song L, Rivera-Pérez JA, Rando OJ. Small RNAs Gained during Epididymal Transit of Sperm Are Essential for Embryonic Development in Mice. Dev Cell. 2018;46:470-480.e3.

42. Fortes MRS, DeAtley KL, Lehnert SA, Burns BM, Reverter A, Hawken RJ, et al. Genomic regions associated with fertility traits in male and female cattle: Advances from

microsatellites to high-density chips and beyond. Animal Reproduction Science. 2013;141:1–19.

43. Abdollahi-Arpanahi R, Gianola D, Peñagaricano F. Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. Genetics Selection Evolution. 2020;52:12.

44. Tarantino C, Paolella G, Cozzuto L, Minopoli G, Pastore L, Parisi S, et al. miRNA 34a, 100, and 137 modulate differentiation of mouse embryonic stem cells. FASEB J. 2010;24:3255–63.

45. Cui Y, Li T, Yang D, Li S, Le W. miR-29 regulates Tet1 expression and contributes to early differentiation of mouse ESCs. Oncotarget. 2016;7:64932–41.

46. Keles E, Malama E, Bozukova S, Siuda M, Wyck S, Witschi U, et al. The micro-RNA content of unsorted cryopreserved bovine sperm and its relation to the fertility of sperm after sex-sorting. BMC Genomics. 2021;22:30.

47. Alves MBR, de Arruda RP, De Bem THC, Florez-Rodriguez SA, Sá Filho MF de, Belleannée C, et al. Sperm-borne miR-216b modulates cell proliferation during early embryo development via K-RAS. Sci Rep. 2019;9:10358.

48. Wei H, Zhou B, Zhang F, Tu Y, Hu Y, Zhang B, et al. Profiling and Identification of Small rDNA-Derived RNAs and Their Potential Biological Functions. PLoS One. 2013;8:e56842.

49. Hua M, Liu W, Chen Y, Zhang F, Xu B, Liu S, et al. Identification of small non-coding RNAs as sperm quality biomarkers for in vitro fertilization. Cell Discov. 2019;5:20.

50. Brendolan A, Rosado MM, Carsetti R, Selleri L, Dear TN. Development and function of the mammalian spleen. Bioessays. 2007;29:166–77.

51. Lopes C, Delezoide A-L, Delabar J-M, Rachidi M. BARHL1 homeogene, the human ortholog of the mouse Barhl1 involved in cerebellum development, shows regional and cellular specificities in restricted domains of developing human central nervous system. Biochemical and Biophysical Research Communications. 2006;339:296–304.

52. Mullen RD, Colvin SC, Hunter CS, Savage JJ, Walvoord EC, Bhangoo APS, et al. Roles of the LHX3 and LHX4 LIM-Homeodomain Factors in Pituitary Development. Mol Cell Endocrinol. 2007;265–266:190–5.

53. Boras-Granic K, Grosschedl R, Hamel PA. Genetic interaction between Lef1 and Alx4 is required for early embryonic development. Int J Dev Biol. 2006;50:601–10.

54. Panman L, Drenth T, Tewelscher P, Zuniga A, Zeller R. Genetic interaction of Gli3 and Alx4 during limb development. Int J Dev Biol. 2005;49:443–8.

55. Qian Y, Fritzsch B, Shirasawa S, Chen C-L, Choi Y, Ma Q. Formation of brainstem (nor)adrenergic centers and first-order relay visceral sensory neurons is dependent on homeodomain protein Rnx/Tlx3. Genes Dev. 2001;15:2533–45.

56. Munroe RJ, Prabhu V, Acland GM, Johnson KR, Harris BS, O'Brien TP, et al. Mouse H6 Homeobox 1 (Hmx1) mutations cause cranial abnormalities and reduced body mass. BMC Dev Biol. 2009;9:27.

57. Ito Y, Toriuchi N, Yoshitaka T, Ueno-Kudoh H, Sato T, Yokoyama S, et al. The Mohawk homeobox gene is a critical regulator of tendon differentiation. Proc Natl Acad Sci U S A. 2010;107:10538–42.

58. Roux M, Laforest B, Eudes N, Bertrand N, Stefanovic S, Zaffran S. Hoxa1 and Hoxb1 are required for pharyngeal arch artery development. Mech Dev. 2017;143:1–8.

59. Xiong X, Tu S, Wang J, Luo S, Yan X. CXXC5: A novel regulator and coordinator of TGF-β, BMP and Wnt signaling. J Cell Mol Med. 2019;23:740–9.

60. Carlone DL, Skalnik DG. CpG binding protein is crucial for early embryonic development. Mol Cell Biol. 2001;21:7601–6.

61. Campbell CE, Piper M, Plachez C, Yeh Y-T, Baizer JS, Osinski JM, et al. The transcription factor Nfix is essential for normal brain development. BMC Dev Biol. 2008;8:52.

62. Landin-Malt A, Benhaddou A, Zider A, Flagiello D. An evolutionary, structural and functional overview of the mammalian TEAD1 and TEAD2 transcription factors. Gene. 2016;591:292–303.

63. Bergemann AD, Cheng HJ, Brambilla R, Klein R, Flanagan JG. ELF-2, a new member of the Eph ligand family, is segmentally expressed in mouse embryos in the region of the hindbrain and newly forming somites. Mol Cell Biol. 1995;15:4921–9.

FIGURE LEGENDS

Figure 1. Data filtering strategy. The four different tables included a heterogeneous number of features **(A).** Because the CpGs, SNPs and sncRNAs were huge data tables, features that could be considered as noise and feature that did not display significant variation among the bulls were filtered out **(B)**. Because the remaining sncRNAs and the SPs were impacted by the extraction batch of the semen, they were next corrected from this batch effect **(C)**. Finally, because the CpG, SNP and sncRNA tables still included an important number of features, the most relevant features were selected using a supervised method: the Random Forest **(D)**. At the end of these three filtering steps, 12,006 relevant features originating from four data tables were kept for further analysis.

Figure 2. Multiple factor analysis highlights the contribution of SNPs, CpGs, and sncRNAs to bull fertility A MFA was run on the 12,006 selected features belonging to the CpG, sncRNA, SNP and SP tables that actively contributed to the results. Furthermore, fertility, bulls' origin and semen extraction batch were set as illustrative features, meaning that they did

not participate in the MFA construction. **A:** Global variable plot with the active features shown in red and the illustrative features in green. **B:** Individual factor map where each dot corresponds to a bull and was coloured depending of its fertility class. C, D: variable factor maps for quantitative features (CpGs and sncRNAs). The first and the second dimensions (C) and the first and the third dimensions (D) are represented. Each arrowhead corresponds to a feature and was coloured depending of its dataset of origin, with CpGs, sncRNAs and SPs shown in blue, yellow and grey, respectively. Furthermore, the intensity of the arrowheads' colour indicated the cos^2, which reflects the strength of the correlation between a feature and dimension 1. In C, two clusters are represented, which gather the features with the most important positive (>0.55, cluster 1) or negative (<0.4, cluster 2) coordinates along dimension 1.

Figure 3. Correlation structure among CpG and sncRNA features. The correlation matrix of the features belonging to the two clusters that were defined Figure 2C was computed. Features are displayed in lines and columns and are coloured according to the datasets and clusters. The intensity of the colour in the heatmap reflects the strength of the correlation between two features, with positive and negative correlations indicated in red and blue, respectively.

Figure 4. Parameter calibration for four methods. A: The four methods used to predict bull fertility. **B:** For Random Forest, the recommended parameters were used, while for other methods parameters were optimised using cross validation (logistic lasso) or one training and one testing set (gradient boosting and neural networks). **C:** for each method that has to be optimized, a grid was set for the main parameters. **D:** For each method, each combination of parameters was tested, and the one maximising the model performance was retained for further analysis.

Figure 5. Prediction accuracy of bull fertility and optimal number of features for the different methods. For each method, one model was constructed with the 12,006 features and features were classified depending on their importance. Then, for each method models

were constructed with the top 1000, 750, 500, 250, 100, 50 and 10 features. Using this information, the figure on the left hand indicates the AUC on the y-axis and the number of features used during model construction on the x-axis. Each dot (coloured according to the method) represents the actual AUC value obtained for each model. A tendency curve was also drawn for each method using the geom_smooth function of the ggplot2 package with default parameters. The table on the right hand shows the optimal number of features and the associated AUC value obtained for each method, based on the actual AUC values and not on the tendency curve.

Figure 6. The selected features are specific to each method. A: Venn diagram showing the intersection between methods in terms of selected features. Areas were coloured according to the proportion of features they included compared to the total amount of features selected by each method. **B:** The datasets of origin of the selected features are represented by pie charts. The methods showing similar behaviours were grouped together.

Figure 7. Functional analyses of the selected features. A: Global strategy for functional analysis. The union of the SNP, CpG and sncRNA features selected by the three unbiased methods (cforest, Gradient Boosting, Neural Networks) was considered and referred to as "Selected features". Genes including the selected CpGs and SNPs were directly subjected to an enrichment analysis. The distribution of different families of sncRNAs highlighted an overrepresentation of miRNAs and rRFs among the selected features when compared to the background, which included the 413,952 sncRNAs that were initially represented in the sncRNA dataset (lower left panel). The analysis was therefore focused on the miRNA target genes that were subjected to a functional enrichment analysis. **B**: The genes containing selected SNP and CpG features were submitted to an enrichment analysis using DAVID. Three clusters of terms were significantly enriched (EASE score above 1.3; left panel). The proportion of genes targeted by selected CpGs only, selected SNPs only, or by both CpGs and SNPs varied in the three clusters (pie charts, right panel). **C**: The genes identified as putative targets of selected miRNAs by Targetscan were submitted to an overrepresentation analysis using

Webgestalt. The top 10 overrepresented GO terms are listed, with the corresponding adjusted p-value.

AVAILABILITY OF DATA AND MATERIALS: additional data files are provided (see below). RRBS and sncRNA fastq files have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession numbers PRJEB46371 (https://www.ebi.ac.uk/ena/data/view/PRJEB46371) and , respectively. The genotypes are owned by the breeding companies and are private data.

ACKNOWLEDGEMENT: this research was possible thanks to the semen samples provide by the two companie Umotest and Evajura. We are also grateful to the Genotoul bioinformatics platform Toulouse Midi-Pyrenees (Bioinfo Genotoul) for providing computing and storage resources. We also want thank Denis Laloe for his help in statistical analysis, Luc Jouneau and Anne Aubert-Frambourg for them help in bioinformatics analysis and Victoria Hawken for English editing.

FUNDING: this study was founded by the French National Research Agency (grant ANR-13-LAB3-0008-01 SeQuaMol) and APIS-GENE (AP-2018-44). VC was a CIFRE fellow of the French National Agency for Research and Technology (ANRT). The funding bodies had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript. APIS-GENE has read and approved the final version of the manuscript.

AUTHOR INFORMATION

Affiliations

[1]Université Paris-Saclay, UVSQ, INRAE, BREED, 78350 Jouy-en-Josas, France.

[2]Ecole Nationale Vétérinaire d'Alfort, BREED, 94700, Maisons-Alfort, France.

[3]R&D Department, ELIANCE, 149 rue de Bercy, 75012, Paris, France.

[4]Université Paris-Saclay, AgroParisTech, INRAE, GABI, 78350 Jouy-en-Josas, France.

[5]INRAE, MaIAGE, Université Paris-Saclay, 78350 Jouy-en-Josas, France.

Contributions: VC analysed data and drafted the manuscript. ES, AB and LS performed the experiments. ES, SM and LS constructed and used the bioinformatics pipeline for the sncRNAs analysis. ES and CH were involved in the experimental design. CH was involved in the analysis of genotypes. LS, ES, HK and FJ conceived the study. LS, HK and FJ obtain the fundings. HK and FJ supervised the study and edited the manuscript. All authors have read and approved the final version of the manuscript.

Corresponding author: Florence Jaffrezic

ETHIC DECLARATIONS

Ethic approval and consent to participate: not applicable (only commercial semen samples used for the purposes of this study).

Consent for publications: not applicable

ADDITIONAL INFORMATIONS

Supplementary informations

Figure 1

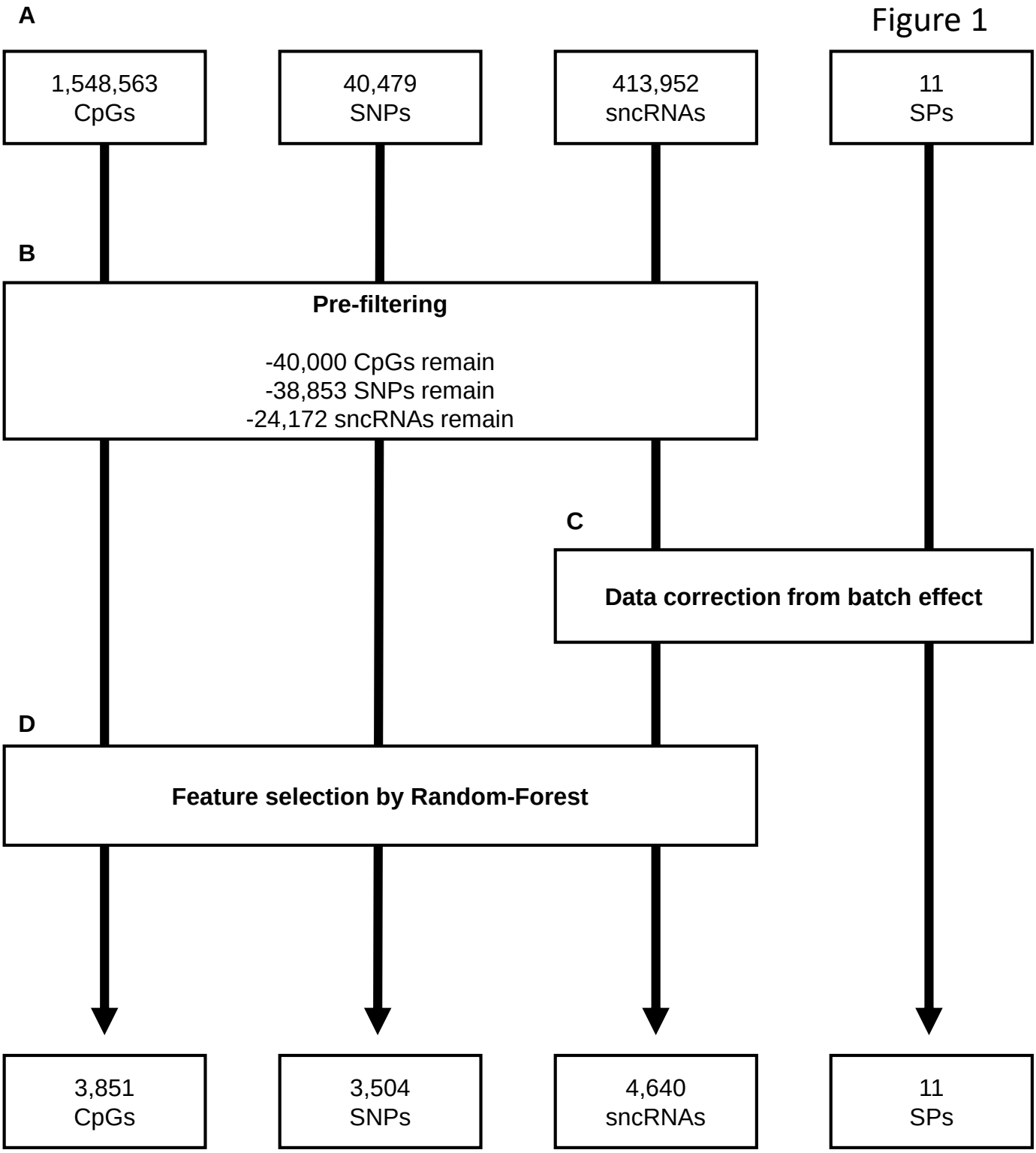
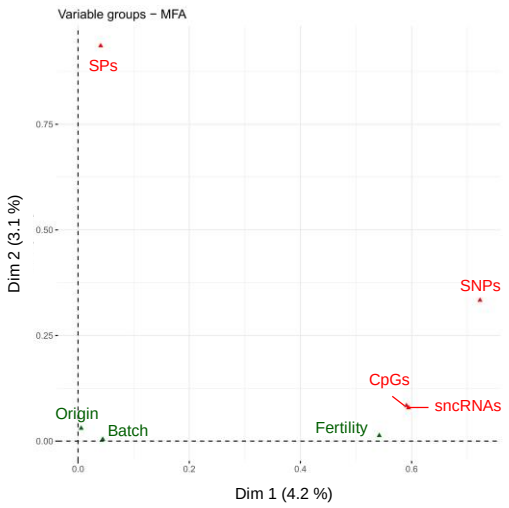
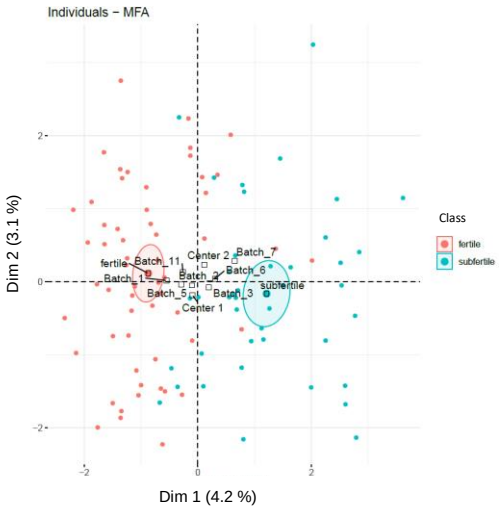

A
1,548,563
CpGs
40,479
SNPs
413,952
sncRNAs
11
SPs
B
Pre-filtering
-40,000 CpGs remain
-38,853 SNPs remain
-24,172 sncRNAs remain
C
Data correction from batch effect
D
Feature selection by Random-Forest
3,851
CpGs
3,504
SNPs
4,640
sncRNAs
11
SPs

Figure 2

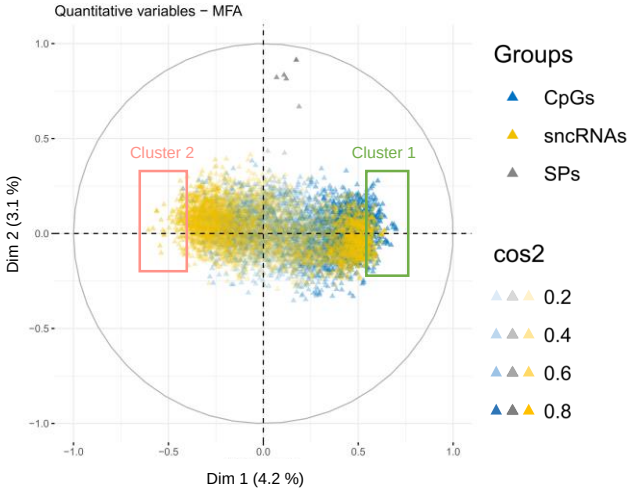
A



B



C



D

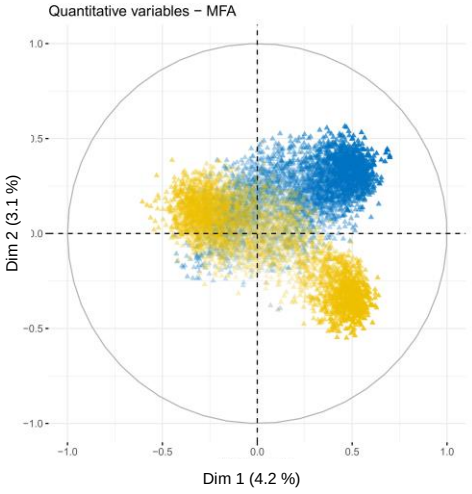


Figure 3

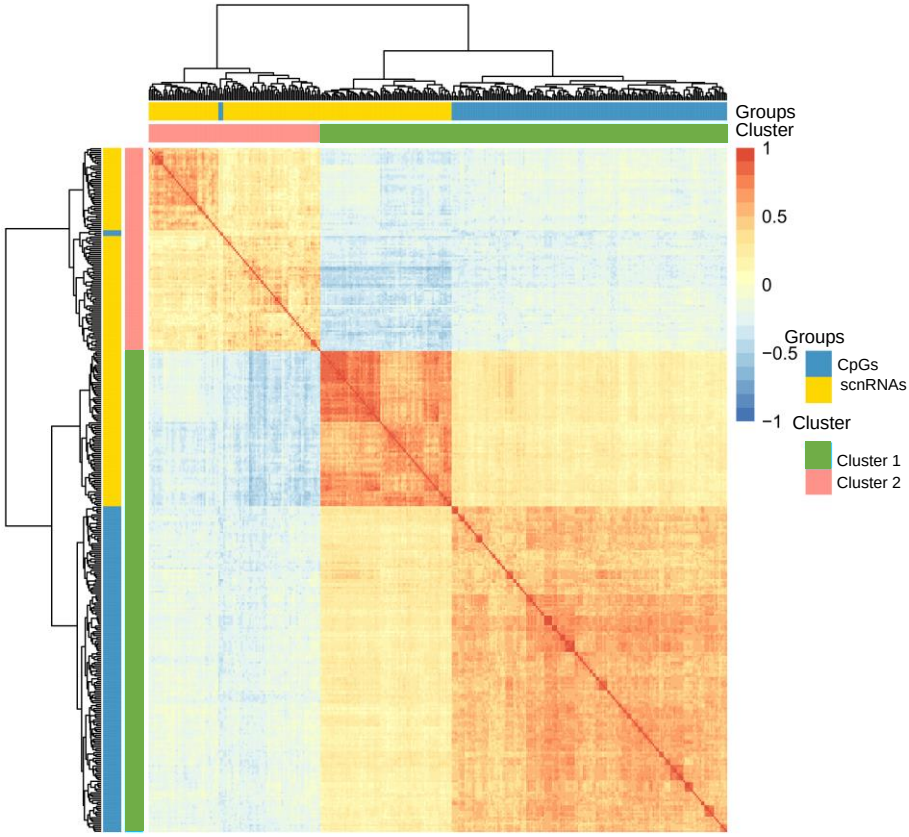

Groups
Cluster
1
0.5
0
−0.5
−1
Groups
CpGs
scnRNAs
Cluster
Cluster 1
Cluster 2

Figure 4

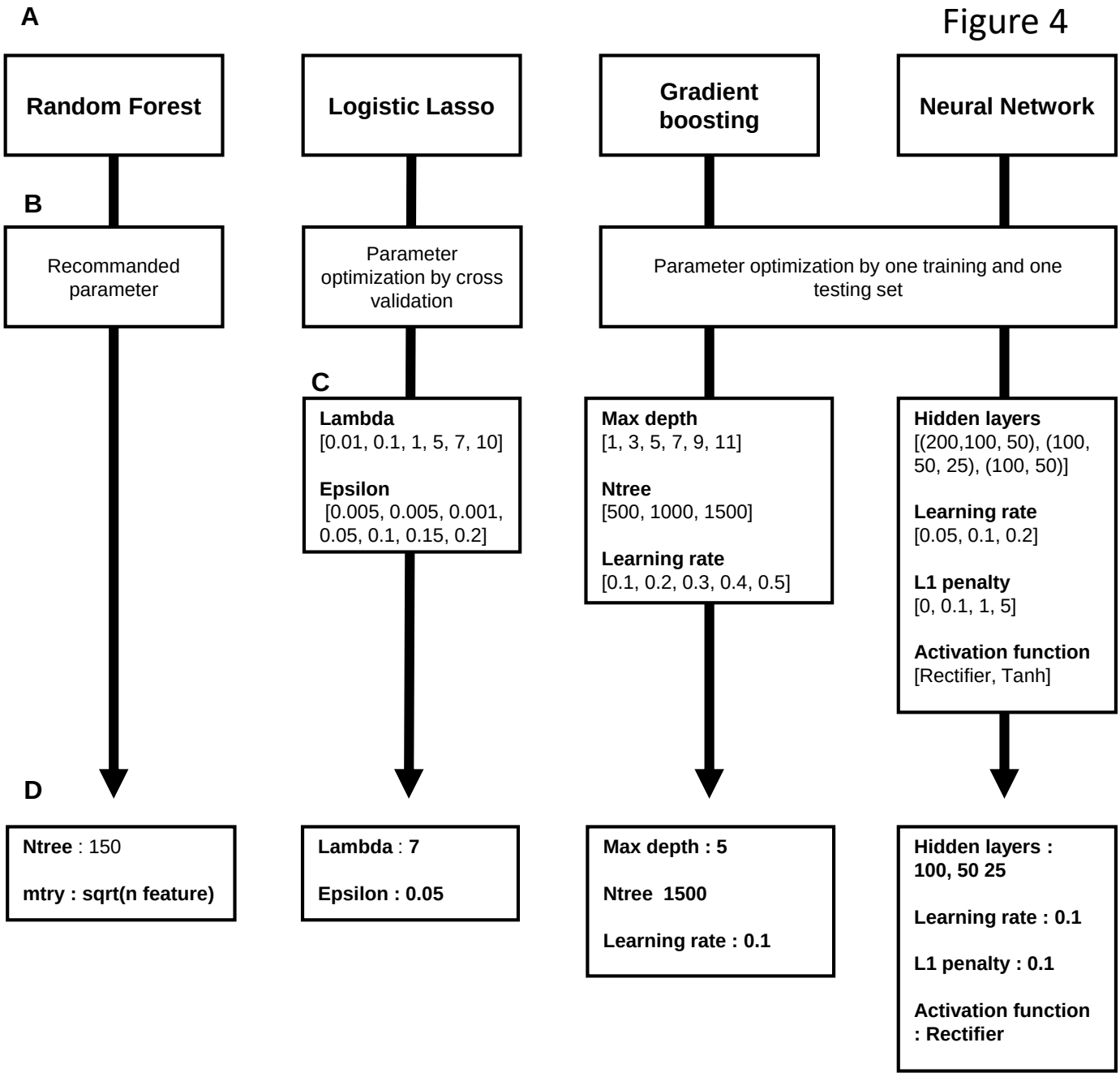
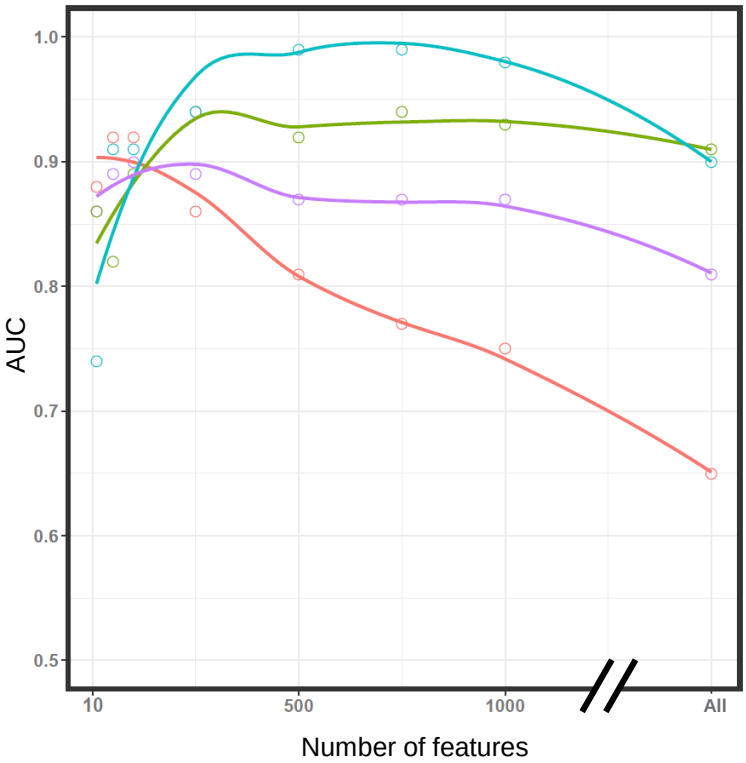

A
Random Forest
Logistic Lasso
Gradient boosting
Neural Network
B
Recommanded parameter
Parameter optimization by cross validation
Parameter optimization by one training and one testing set
C
Lambda
[0.01, 0.1, 1, 5, 7, 10]
Epsilon
[0.005, 0.005, 0.001, 0.05, 0.1, 0.15, 0.2]
Max depth
[1, 3, 5, 7, 9, 11]
Ntree
[500, 1000, 1500]
Learning rate
[0.1, 0.2, 0.3, 0.4, 0.5]
Hidden layers
[(200,100, 50), (100, 50, 25), (100, 50)]
Learning rate
[0.05, 0.1, 0.2]
L1 penalty
[0, 0.1, 1, 5]
Activation function
[Rectifier, Tanh]
D
Ntree : 150
mtry : sqrt(n feature)
Lambda : 7
Epsilon : 0.05
Max depth : 5
Ntree 1500
Learning rate : 0.1
Hidden layers : 100, 50 25
Learning rate : 0.1
L1 penalty : 0.1
Activation function : Rectifier

Figure 5

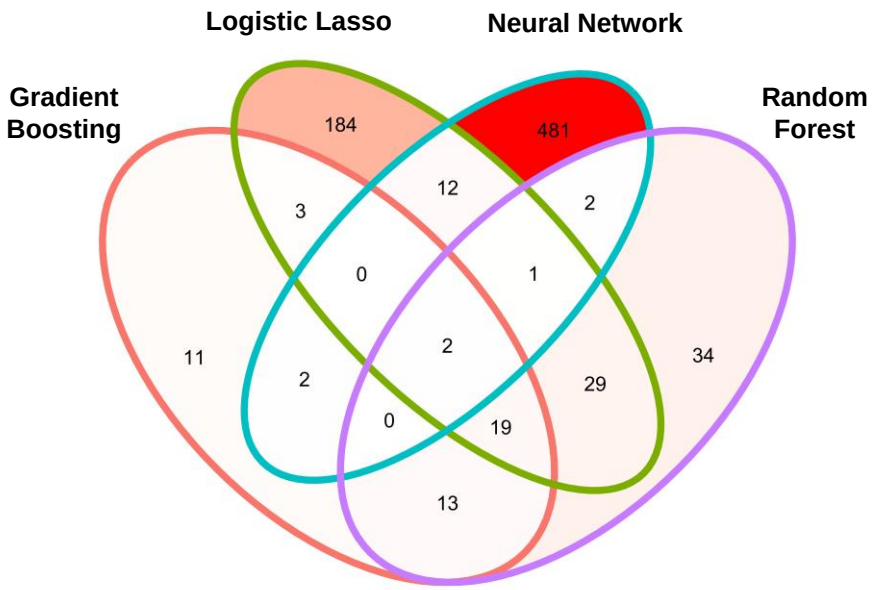


	Number of optimal features	AUC
Random Forest	100	0.9
Gradient Boosting	50	0.92
Neural Network	500	0.99
Logistic Lasso	250	0.94

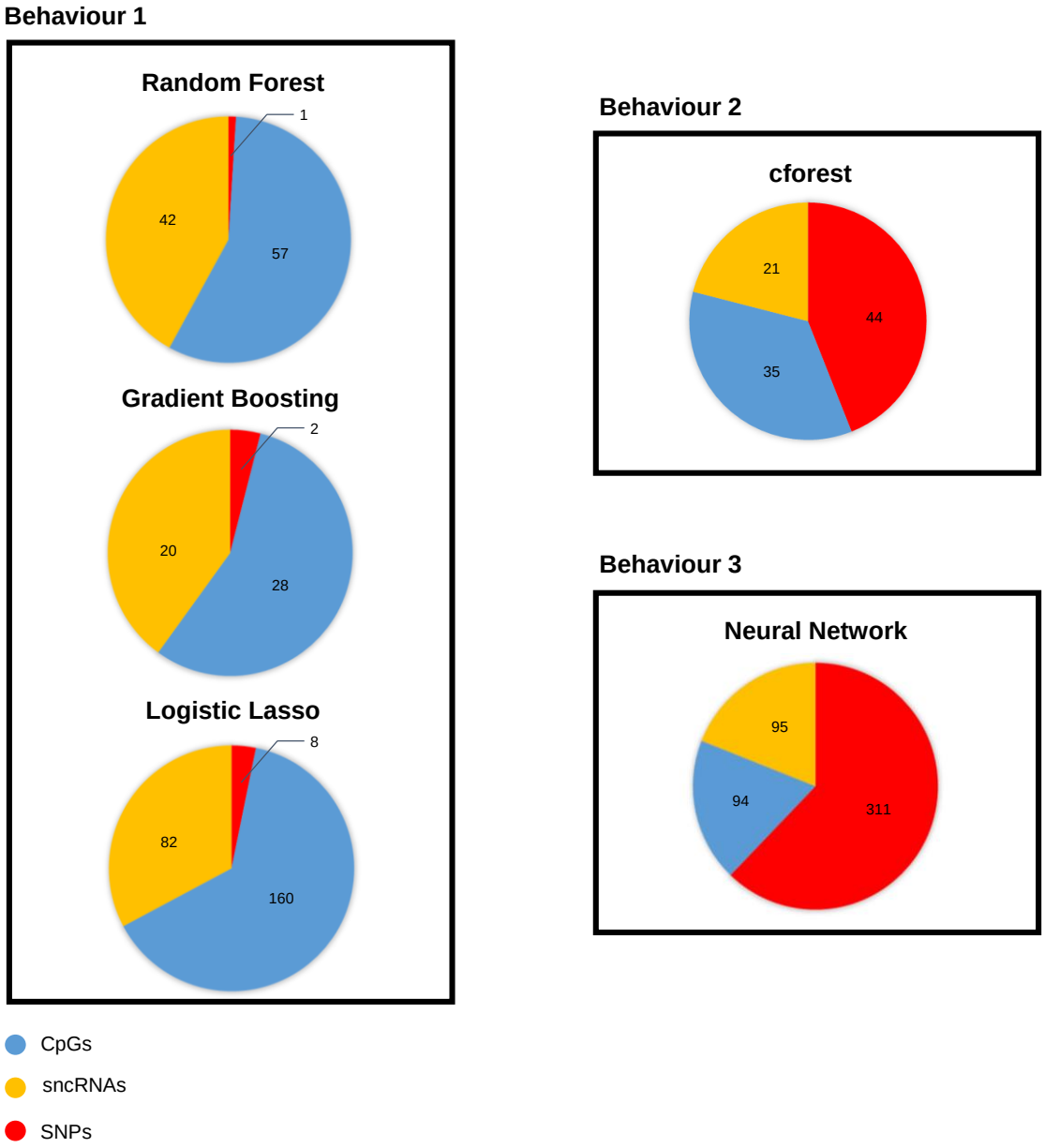
- Random Forest
- Gradient Boosting
- Neural Network
- Logistic Lasso

Figure 6

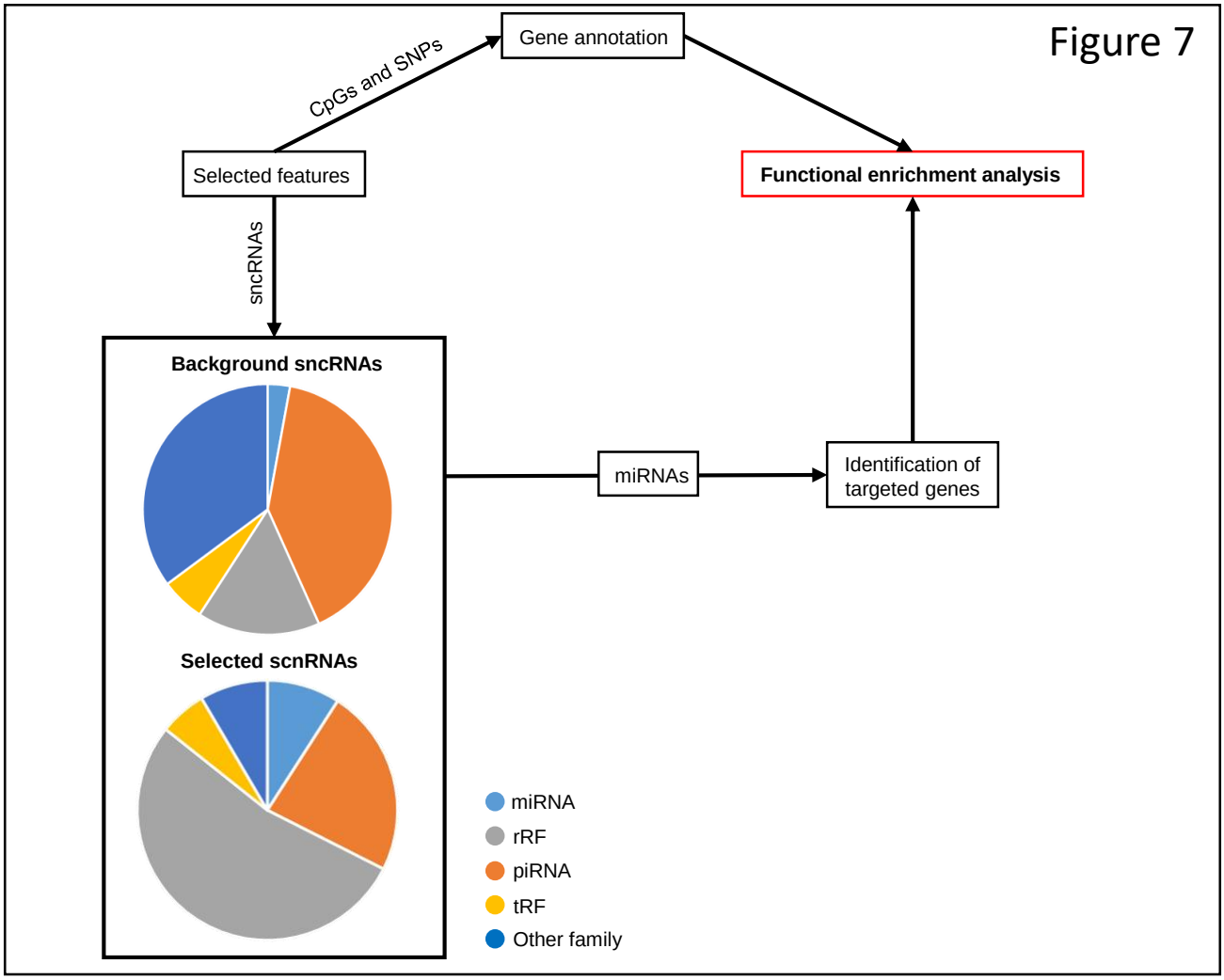
A


Logistic Lasso
Neural Network
Gradient Boosting
Random Forest
184
481
12
3
2
0
1
11
2
2
34
29
0
19
13

B

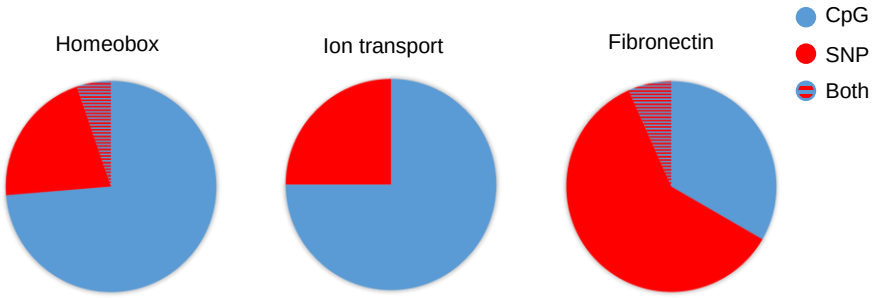

Behaviour 1
Random Forest
1
42
57
Gradient Boosting
2
20
28
Logistic Lasso
8
82
160
Behaviour 2
cforest
21
44
35
Behaviour 3
Neural Network
95
94
311
CpGs
sncRNAs
SNPs

A



B

Cluster	EASE score
Homeobox	1.77
Ion transport	1.6
Fibronectin	1.42



C

Overrepresented terms among miRNA target genes	Adjusted p-value (FDR)
Regulation of cell morphogenesis (GO:0022604)	3.2^{e}-10
Response to transforming growth factor beta (GO:0071559)	3.2^{e}-10
Morphogenesis of an epithelium (GO:0002009)	9.3^{e}-10
Synapse organization (GO:0050808)	3.3^{e}-9
Axon development (GO:0061564)	3.3^{e}-9
Cell-cell signaling by Wnt (GO:0198738)	4.6^{e}-9
Transmembrane receptor protein STK pathway (GO:0007178)	4.6^{e}-9
Regulation of neuron projection development (GO:0010975)	9.7^{e}-9
Morphogenesis of a branching structure (GO:0001763)	1.8^{e}-9
Organ growth (GO:0035265)	2.2^{e}-8

II.I.II : Résultats préliminaires en inférence de réseaux

Problématique

Dans l'article d'intégration de données en race Montbéliarde, un des objectifs était d'analyser les liens pouvant exister entre les différents types de variables. Cela a été réalisé en sélectionnant les variables les mieux représentées par le premier axe de l'AFM. Nous avions pu observer peu de corrélations entre les sncRNAs et CpG. Néanmoins, l'AFM ne permet pas de sélectionner des variables quantitatives et qualitative en même temps, ainsi les SNP n'avaient pas été examinés.

Or, ce type de données est important dans la construction des modèles de prédiction de la fertilité, au même titre que les CpG et les sncRNA. Ainsi, analyser l'ensemble de ces informations pourrait être intéressant. Pour cela, nous avons décidé d'utiliser les méthodes d'inférence de réseaux. Comme décrit dans la partie Introduction, l'inférence de réseaux dans le cadre des données -omiques peut être difficile car il y a souvent plus de variables que d'individus, et que les données ne sont pas forcément de même nature. Dans notre cas, nous disposons de deux types de données quantitatives (sncRNA et CpG) n'étant pas toutes gaussiennes et d'un type de variables qualitatives à trois modalités (SNP). Nous avons donc décidé d'utiliser la méthode développée dans le package R GENIE3, se basant sur des forêts aléatoires. Cette méthode est peu sensible au nombre et aux types de données en entrée. Les travaux présentés dans cette partie sont des résultats préliminaires, qui seront évidemment à approfondir.

Méthode

<u>Type de données utilisées</u>

Ce travail a été initié en sélectionnant un sous-échantillon des variables les plus pertinentes identifiées par les méthodes cforest, Réseaux de Neurones et Logistic Lasso. Ce sous-échantillon a été constitué en faisant un compromis entre importance des variables dans les modèles, lien des variables avec des gènes impliqués dans la fertilité et représentation équitable des variables entre types de données.

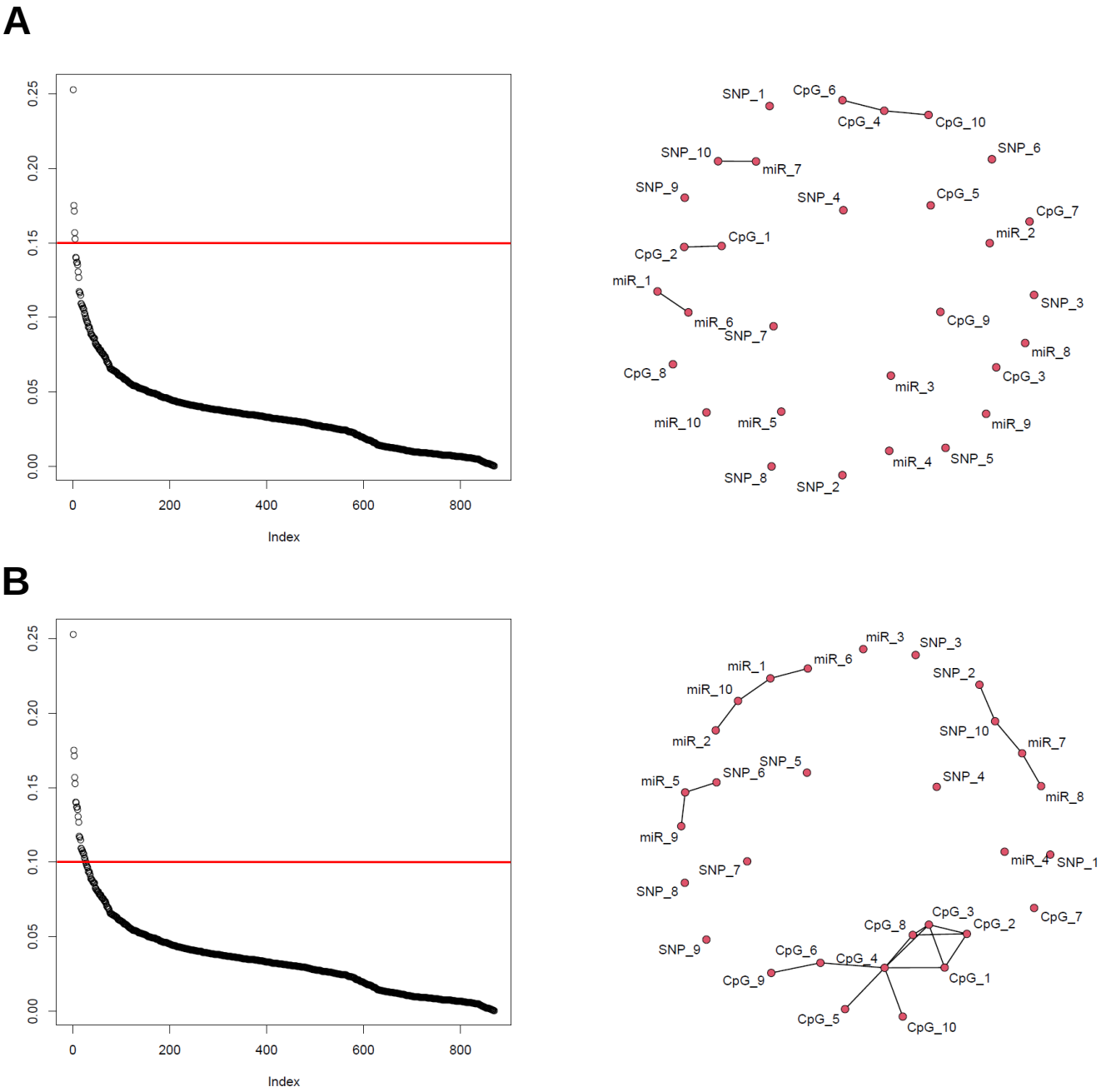


Figure 36 : Inférence de réseaux, montrant que les variables interagissent principalement au sein d'une seule couche d'-omiques. Deux seuils arbitraires ont été choisis par l'utilisateur, 0.15 (A) et 0.1 (B). Les réseaux associés construits à partir du package R network sont présentés sur la droite de la figure. Chaque variable est représentée par un point, et les interactions identifiées sont représentées par des arcs entre variables. Les importances relatives des interactions obtenues par le package R GENIE3 ont été classées par ordre décroissant et représentées sur la gauche de la figure.

L'objectif était de construire un jeu de données de 30 variables, représenté par 10 variables de chaque type.

Dans un premier temps, nous avons voulu identifier les variables avec la plus grande importance dans les différents modèles. Pour chaque variable, les scores d'importance obtenus par les trois méthodes ont été centrés et réduits afin d'être comparables d'un modèle à l'autre. La somme de ces scores a été calculée, permettant de classer les variables avec un score unique, reflétant son importance sur les différents modèles.

Dans un second temps, nous avons voulu sélectionner en priorité les variables associées à des gènes liés à des fonctions biologiques pertinentes. Pour les SNP cette démarche a conduit à la sélection de 10 variables associées aux gènes : *AURKC, CRIM1, CDX1, CDH18, SS18, FCRL3, PTPRG, WIZ, DLG et LSM6*. Pour les CpG les 10 variables sont associées aux gènes : *NANOS2, CXXC1, NFIX, ZFHX4, DMRT3, CHD6, YTHDC2, TEAD2, LHX3 et BARHL1*. Pour les sncRNAs, nous nous sommes concentrés sur les miRNA et avons sélectionné miR-100, miR-2483, miR-296, miR-2285, miR-34, miR-339, miR-30, miR-19 et miR-21.

Méthodes utilisées

Pour réaliser ce travail nous nous sommes basés sur le package GENIE3 qui se base sur les forêts aléatoires dans le but d'estimer les liens entre variables, et avons utilisé les paramètres recommandés par l'auteur : 1000 arbres de décision et un mtry égal au nombre de variables dans le modèle, autrement dit, sans sélection aléatoire de variables.

Cependant, les travaux décrits dans l'article 2 ont montré, conformément à ce qui est indiqué dans la littérature, que les forêts aléatoires étaient biaisées pour la sélection de variables quantitatives au détriment des variables qualitatives (Strobl *et al.*, 2007). Ainsi, en nous inspirant de la méthode développée dans GENIE3, nous avons implémenté la même stratégie, mais cette fois-ci en utilisant le cforest. Le package R « Network » avec le paramètre « directed = FALSE » (qui permet de faire des

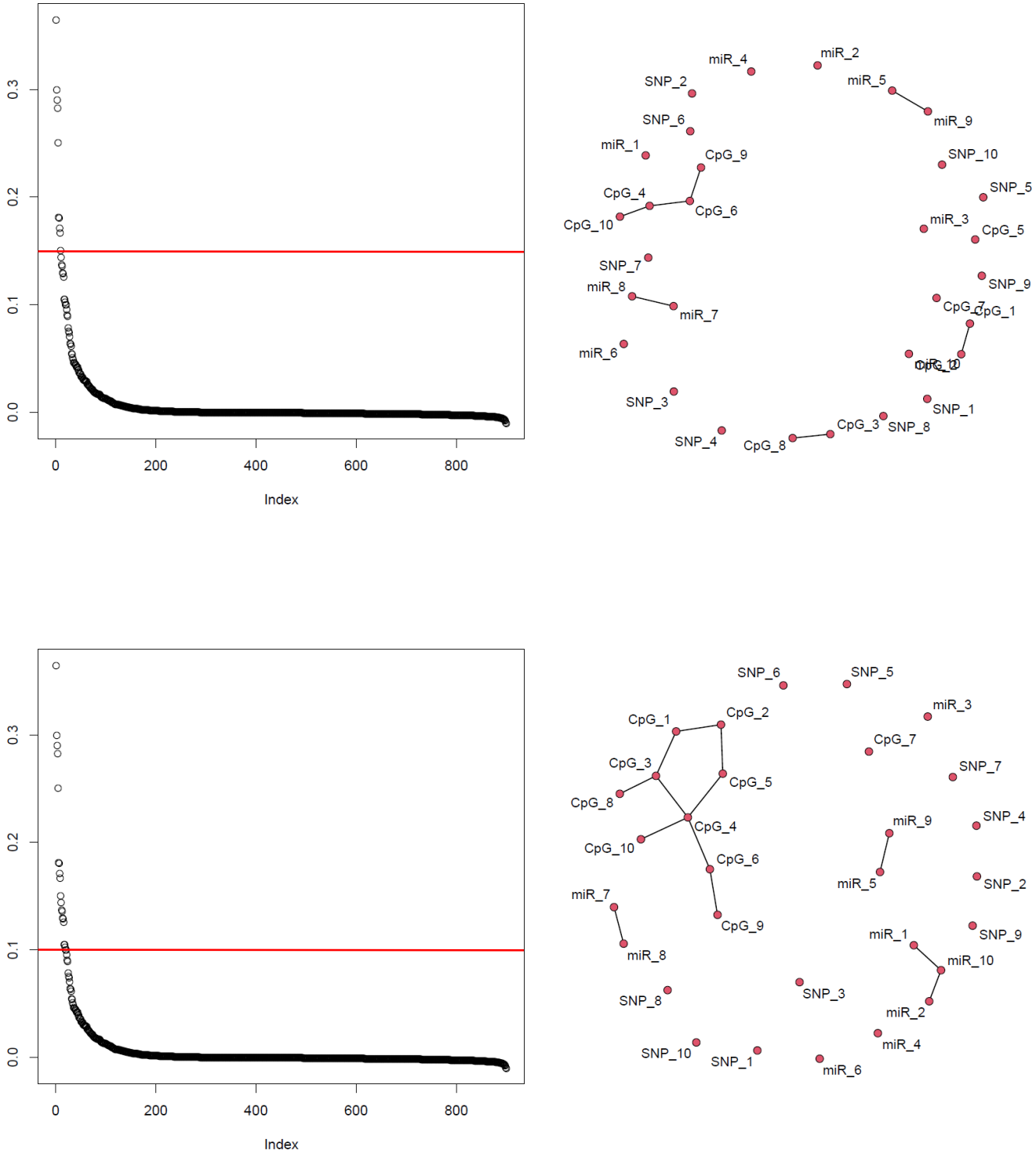


Figure 37 : Les inférences de réseaux basé sur la méthode des cforest n'identifient que des interactions au sein d'une couche unique d'-omiques. Deux seuils arbitraires ont été choisis par l'utilisateur, 0.15 (A) et 0.1 (B). Les réseaux associés construits à partir du package R network sont présentés sur la droite de la figure. Chaque variable est représentée par un point, et les interactions identifiées sont représentées par des arcs entre variables. Les importances relatives des interactions obtenues par le package cforest ont été classées par ordre décroissant et représentées sur la gauche de la figure.

réseaux sans flèches sur les arcs), et tous les autres paramètres par défaut, a été utilisé pour visualiser les réseaux obtenus.

Résultats

Résultats obtenus avec le package GENIE3

Dans un premier temps, nous avons réalisé l'inférence en nous basant sur la méthode de GENIE3. Une des difficultés de cette approche, c'est qu'il n'y a pas de façon (en tout cas implémentée dans le package) permettant de définir un seuil de significativité à partir duquel une interaction entre variables peut être considérée comme significative. Ainsi, nous avons considéré différents niveaux de seuils et les résultats sont présentés Figure 36. Les seuils ont été définis manuellement de sorte que les interactions entre variables ayant une valeur inférieure soient fixées à 0 et n'apparaissent pas dans le réseau. Deux seuils ont été définis : une importance relative strictement supérieure à 0.15 puis à 0.1.

A 0.15 qui est un seuil drastique, le nombre d'arcs présents entre variables est faible (5 arcs) (Figure 36A). On remarque que la plupart de ces arcs (4 parmi 5), lient des variables au sein d'une seule couche d'-omiques. Nous avons donc abaissé le seuil à 0.1, ce qui a permis de construire des réseaux avec plus d'arcs (20). Parmi ces 20 arcs, il n'y en a que deux qui lient deux couches d'-omiques différentes, mettant donc en lumière une majorité d'interaction des variables au sein d'une même couche de données –omiques. On observe en particulier un grand réseau composé de 9 parmi les 10 CpG inclus dans la construction de ces réseaux, suggérant une structure de corrélation entre ces différents CpG.

Résultats obtenus en implémentant la méthode « cforest »

Comme expliqué précédemment les méthodes des forêts aléatoires sont biaisées dans leur sélection. Dans le but d'examiner si ce biais pourrait interférer avec l'identification des interactions, privilégiant notamment les interactions mono-omiques, nous avons implémenté GENIE3 en utilisant le cforest à la place des forêts aléatoires. Les résultats sont présentés Figure 37. Dans un premier temps, un seuil à 0.15 a été fixé, conduisant encore une fois à des réseaux d'interaction de petites dimensions et tous

de nature mono-omique. En diminuant la stringence du seuil, on retrouve les grands réseaux de CpG identifiés précédemment, ainsi que d'autres arcs, tous de nature mono-omique. Cela suggère que l'identification d'interactions mono-omiques ne sont pas du fait de la technique utilisée, mais plutôt d'une structure d'indépendance caractérisant le jeu de données entre les différents types de données omiques.

Discussion

Dans ces résultats préliminaires sur l'inférence de réseaux nous avons pu observer, dans un premier temps, qu'a des seuils élevés les variables liées entre elles étaient majoritairement de nature mono-omiques, et les réseaux étaient de petite taille. En abaissant les seuils, on observe assez naturellement des réseaux plus grands, avec encore une fois des liaisons principalement mono-omiques. Ces résultats très préliminaires tendent à rejoindre les observations faites par AFM vis-à-vis de la structure d'indépendance des variables. Cependant ces résultats sont à approfondir à plusieurs niveaux.

Tout d'abord dans ce travail nous nous sommes basés sur un jeu de données réduit en faisant un compromis entre importance des variables et rôles biologiques potentiels. Néanmoins cette sélection pose de nombreuses questions, en particulier vis-à-vis du rôle biologique qui s'appuie sur des connaissances *a priori* pas nécessairement pertinentes pour l'espèce bovine. Ces résultats préliminaires obtenus sur une poignée de variables seront donc à étayer avec un jeu de données plus conséquent, s'appuyant sur une stratégie de sélection différente.

Certaines pistes d'amélioration sont à effectuer en particulier vis-à-vis de la significativité des liens identifiés. En effet, la métrique utilisée dans cette méthode est l'importance relative, qui ne témoigne en rien de la significativité d'une relation entre deux variables. Ainsi, des couples de variables un peu liées l'une à l'autre peuvent être mis en lumière, sans pour autant que ces liaisons soient significatives. Fellinghauer et al, ont mis au point une méthodologie de « Stability selection » dans le cadre des inférences de réseaux basées sur la méthodologie des forêts aléatoires (Fellinghauer *et al.*, 2013). Ces

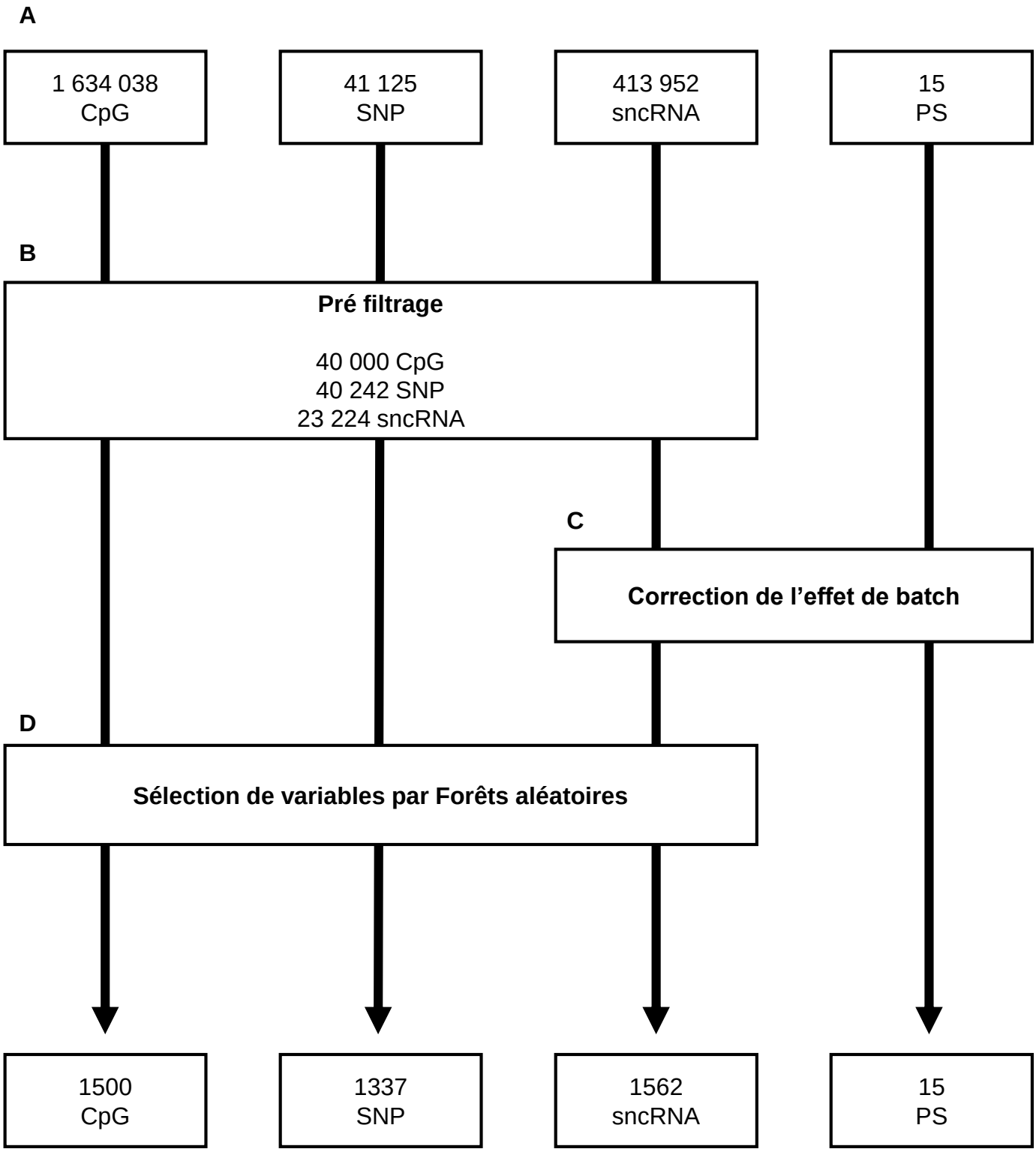


Figure 38 : Pré-traitement des données pour l'intégration de données en race Holstein. A : Les quatre différentes tables de données initiales sont composées d'un nombre hétérogène de variables. B : Parce que les tables de CpG, les sncRNA et les SNP contiennent un grand nombre de données, dont la plupart ne sont pas pertinentes, elles ont été soumises à un pré-filtrage ayant pour objectif de supprimer les données peu variables entre individus ou assimilées à du bruit de fond. C : Un effet batch a été mis en évidence dans les données de paramètres spermatiques (PS) et de sncRNA, qui ont été corrigées de cet effet. D : Les données ont été sélectionnées par rapport à leur lien à la fertilité, à l'aide d'une méthode supervisée par forêts aléatoires.

méthodologies permettent de sélectionner les liens entre variables étant robustes et donc pourront être implémentées dans nos travaux pour tester la significativité des liens identifiés.

Enfin une grande partie du travail qui n'a pour l'instant pas pu être réalisé est le fait de donner un sens biologique aux interactions mises en évidence. Nous avons pu remarquer que la majorité des CpG sélectionnés étaient liés. Il pourrait être intéressant par exemple d'examiner si les niveaux de méthylation de l'ADN au niveau de ces CpG étaient co-régulés par un facteur commun. Il est également possible que ces CpG soient tous liés car localisés dans une même région génomique.

II.II : Intégration de données en race Holstein

Un travail d'intégration de données a également été réalisé dans la race Holstein et les résultats sont présentés ci-dessous. Ces travaux ont été conduits sur la cohorte « fertilité » en race Holstein en analysant le méthylome, les sncRNA, les SNP et les paramètres spermatiques des animaux.

Préparation des tables de données

Les tables de données sont de dimensions hétérogènes avec un background de 1 634 038 CpG (après suppressions des CpG polymorphes), 41 125 SNP ayant passé les filtres de contrôle qualité, 413 952 sncRNA normalisés et 15 paramètres spermatiques (Figure 38A). Les trois tables de plus grande dimension, contenant un grand nombre de variables peu pertinentes, ont dans un premier temps été pré-filtrées (Figure 38B). Ce pré-filtrage a consisté à supprimer les données peu ou pas variables pour les CpG et les SNP, ainsi que les ARN peu exprimés dans la table des sncRNA, selon les mêmes modalités que présenté en race Montbéliarde dans l'article 2. Cela a permis de réduire les tables à 40 000 CpG, 40 242 SNP et 23 224 scnRNA. A la suite d'analyses descriptives, un effet lot d'extraction (effet « batch ») a été mis en évidence pour les tables des sncRNA et des paramètres spermatiques mais par pour les CpG (Figure 38C). De ce fait, ces deux tables ont été corrigées pour cet effet à l'aide d'un modèle linéaire comme cela a été réalisé en race Montbéliarde. Enfin, les tables des CpG, SNP et sncRNA comportaient encore un trop grand nombre de variables non nécessairement liées à la fertilité.

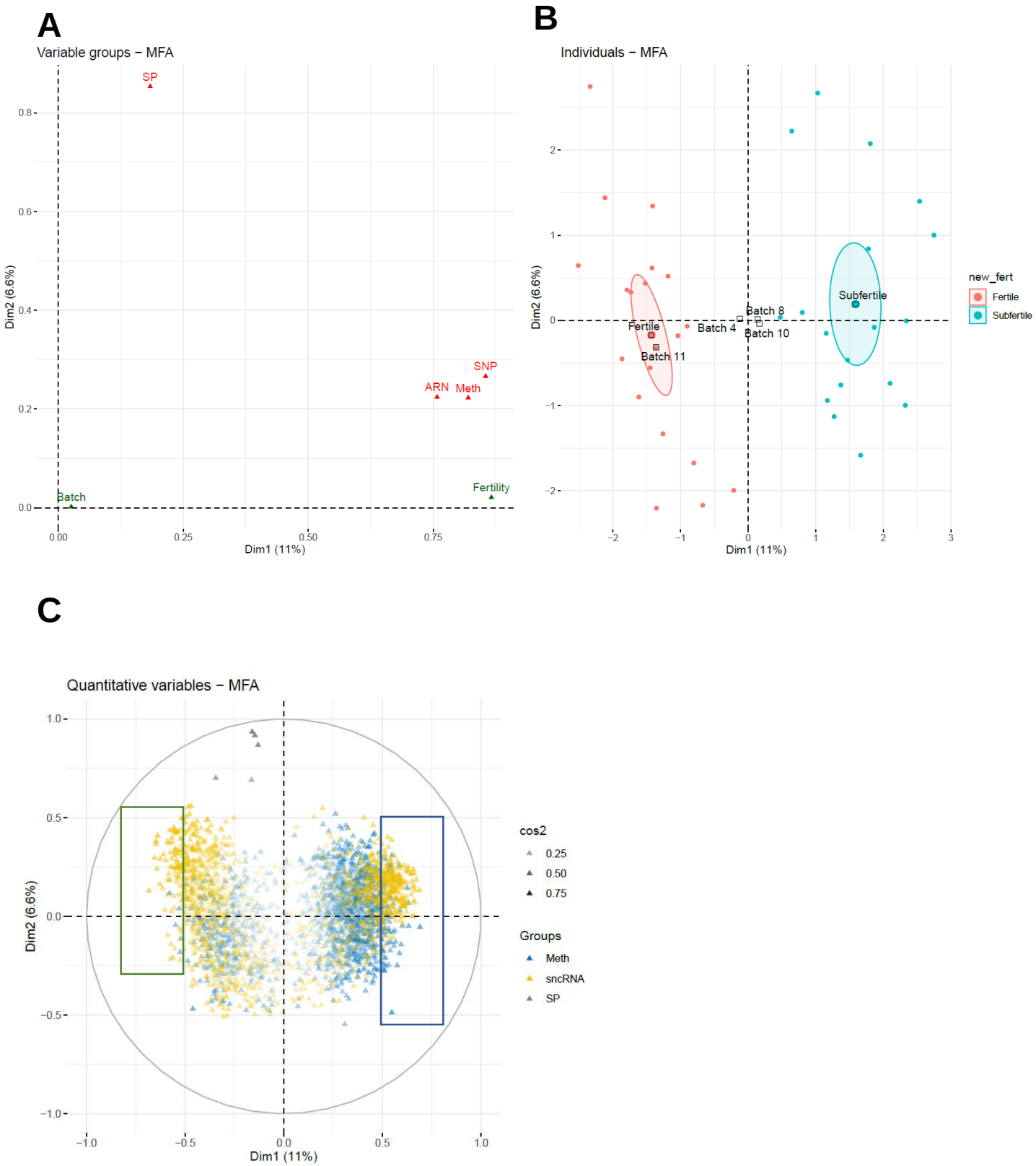


Figure 39 : Résultats de l'AFM en race Holstein. A : Projection des groupes de variables sur le plan factoriel défini par les axes 1 et 2. Les groupes de variables en rouges sont dit actifs et contribue directement à la construction de l'AFM, les variables en vert sont inactives et ont un but illustratif. B : Graphiques des individus sur le plan factoriel défini par les axes 1 et 2. Chaque point représente la projection d'un individu sur ce plan, et est colorié en fonction de la fertilité des animaux. C : Projection des variables quantitatives sur le plan factoriel défini par les axes 1 et 2. Chaque tête de flèche correspond à une variable, et est coloriée en fonction du groupe auquel elle appartient. L'intensité de la couleur dépend du cos^2. Les deux groupes de variables avec des coordonnées extrêmes sont montrés.

C'est pourquoi elles ont été soumises à une sélection de variables par une méthode supervisée –les forêts aléatoires-(Figure 38D). Cette sélection n'ayant pas pour objectif d'être drastique, des seuils faibles sur la contribution des variables aux modèles ont été appliqués (les mêmes qu'en race Montbéliarde). Cela a permis de sélectionner 1500 CpG, 1337 SNP, 1562 sncRNA. A la fin de ce processus de préparation des données, un jeu de 4414 variables représentant quatre types de données différentes a donc été constitué.

Analyses par AFM

Une AFM a été construite en utilisant les CpG, sncRNA, SNP et paramètres spermatiques (PS) comme variables actives et la fertilité et l'effet batch comme variables illustratives. Le graphe des groupes de variables est représenté en Figure 39A. Le premier axe explique 11 % de la variance totale de ces jeux de données et comme en race Montbéliarde on remarque que la fertilité est liée à cet axe. On peut en effet observer sur le graphe des individus (Figure 39B), que les animaux fertiles sont séparés des animaux subfertiles par ce premier axe. On remarque que les types de données liés à cet axe sont les sncRNA, les CpG et les SNP, suggérant leur implication dans la construction de ce dernier et leur liaison à la fertilité. Les paramètres spermatiques, eux, sont comme en Montbéliarde très contributeurs au second axe, mais peu au premier, suggérant leur indépendance vis-à-vis de la fertilité des taureaux. Le graphe des variables quantitatives est présenté Figure 39C. On remarque que les variables appartenant aux sncRNA et CpG ont aussi bien une coordonnée positive sur le premier axe (donc hyperméthylées ou sur-exprimés chez les taureaux subfertiles) qu'une coordonnée négative (hypométhylées ou sous-exprimés chez les taureaux subfertiles). Si les sncRNA surexprimés/sous-exprimés chez les taureaux subfertiles sont présents en quantité équilibrée, ce n'est pas le cas pour les CpG, déséquilibrés en faveur d'une hyperméthylation chez les taureaux subfertiles. Cette observation rejoint les résultats de l'analyse différentielle où une majorité de DMC présentaient la même tendance. On remarque également que les sncRNA sélectionnés sont plus nombreux que les CpG sélectionnés.

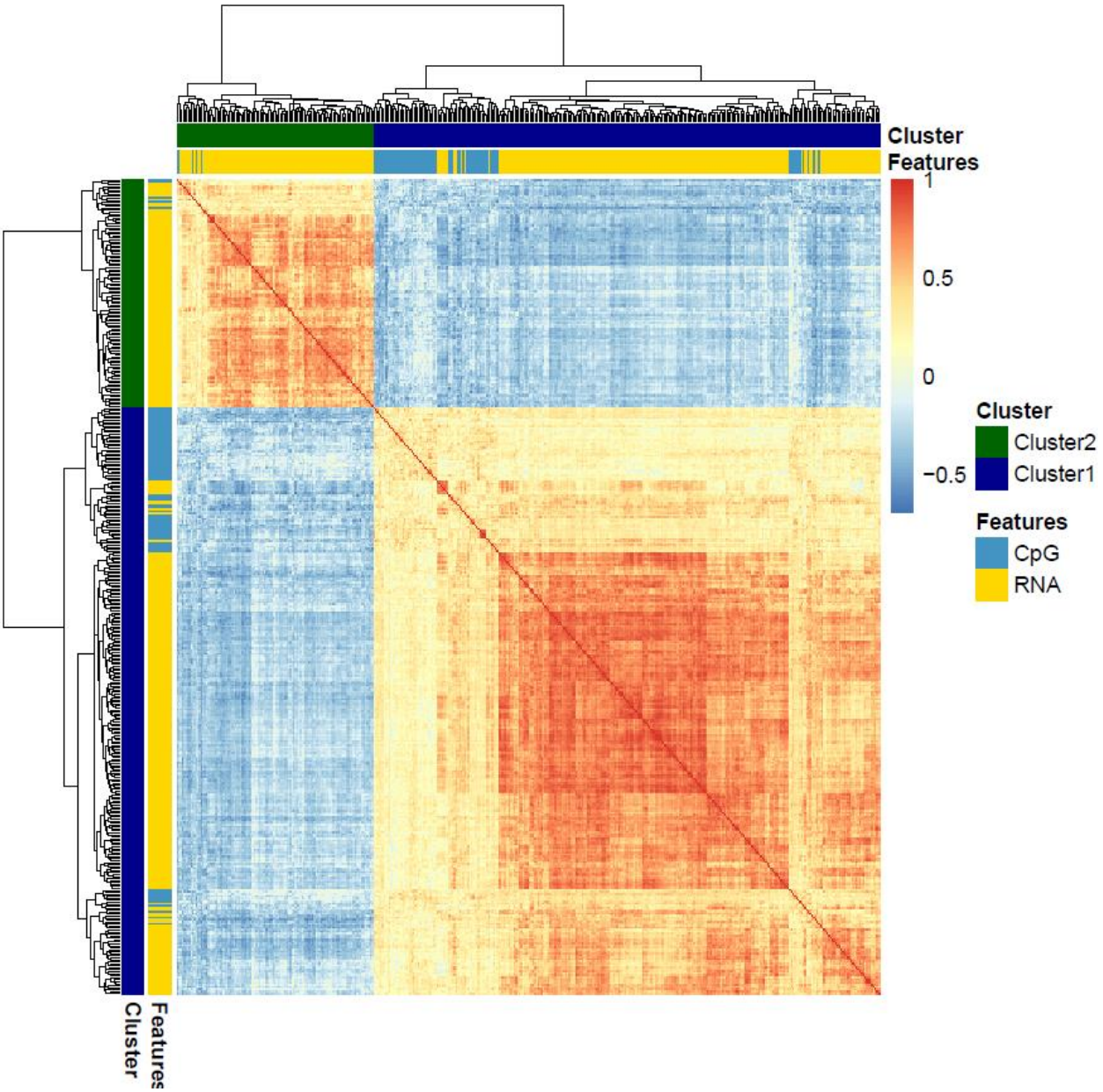


Figure 40 : Heatmap des variables quantitatives les mieux représentées par le premier axe de l'AFM. Une matrice de corrélation de Pearson a été construite en utilisant les variables quantitatives des deux groupes définis dans l'AFM. Cette matrice de corrélation a ensuite été représentée avec une classification à l'aide du package pheatmap disponible dans R.

Corrélation entre méthylation de l'ADN et sncRNA

La fertilité des animaux est liée au premier axe de l'AFM, ce qui signifie que les variables contribuant le plus à cette dernière varient en fonction de la fertilité des animaux. Nous avons donc analysé les variables quantitatives les plus liées à cet axe. Pour cela nous avons appliqué des filtres aux coordonnées de ces variables sur le premier axe. Nous avons donc sélectionné les variables ayant une coordonnée supérieure à 0.5 ou inférieure à -0.5, ce qui a permis de définir deux groupes différents. Les corrélations entre variables ont été visualisées en réalisant une « heatmap » avec classification à partir de la matrice de corrélation des variables de ces deux groupes (Figure 40).

On n'observe pas de corrélations élevées entre les CpG et les sncRNA, mais plutôt au sein d'un seul type de données, avec notamment des corrélations élevées observées pour un noyau de sncRNA du groupe 1. Cela montre que les variables liées à la fertilité par AFM sont peu liées entre elle.

Optimisation des hyper-paramètres des méthodes d'intégration

L'objectif suivant a été de construire des modèles de prédiction dans le but de déterminer s'il était possible d'améliorer encore la prédiction de la fertilité, et de sélectionner les variables les plus prédictives. Avant cela, il faut choisir les hyper-paramètres propres à chacune des méthodes utilisées. Dans ce travail en race Holstein, seules trois méthodes ont été utilisées (Figure 41A), contrairement à la race Montbéliarde où cinq méthodes avaient été évaluées. Dans l'étude en race Montbéliarde nous avions en effet mis en évidence un biais dans la sélection de variables quantitatives avec les méthodes de forêts aléatoires et de gradient boosting, qui n'ont donc pas été réutilisées en race Holstein. Les étapes permettant de définir les hyper-paramètres sont les mêmes qu'en race Montbéliarde (Figure 4 de l'article 2) et ne seront donc pas re-détaillées ici. Néanmoins, le détail complet de ce processus et les paramètres retenus sont détaillés en Figure 41.

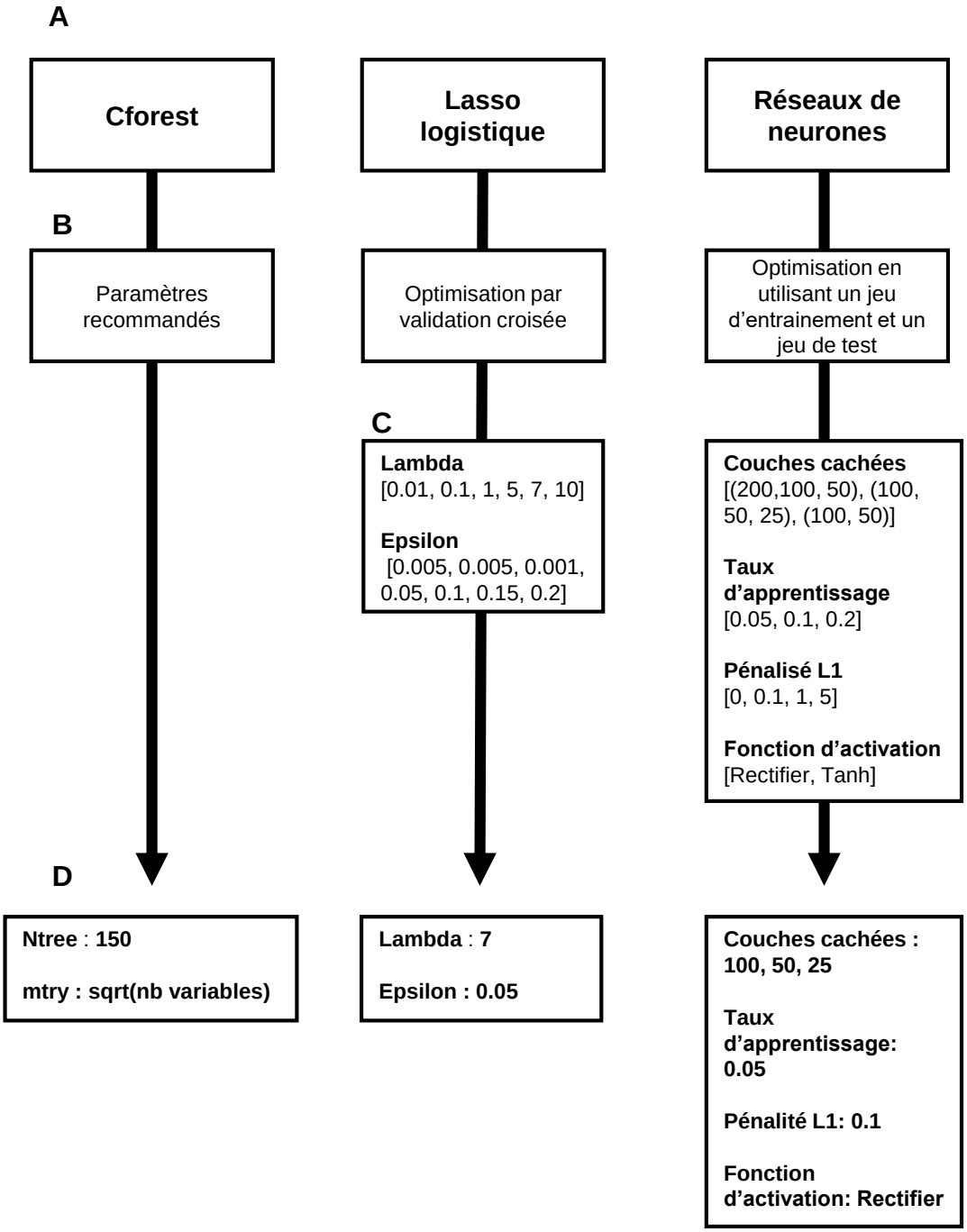


Figure 41 : Présentation de l'optimisation des hyper-paramètres des modèles. A : Méthodes utilisées. B : Type de stratégies utilisées pour l'optimisation des hyper-paramètres des modèles. C : Liste et grille de valeurs des paramètres optimisés pour chaque méthode. D : Hyper-paramètres choisis pour chaque méthode

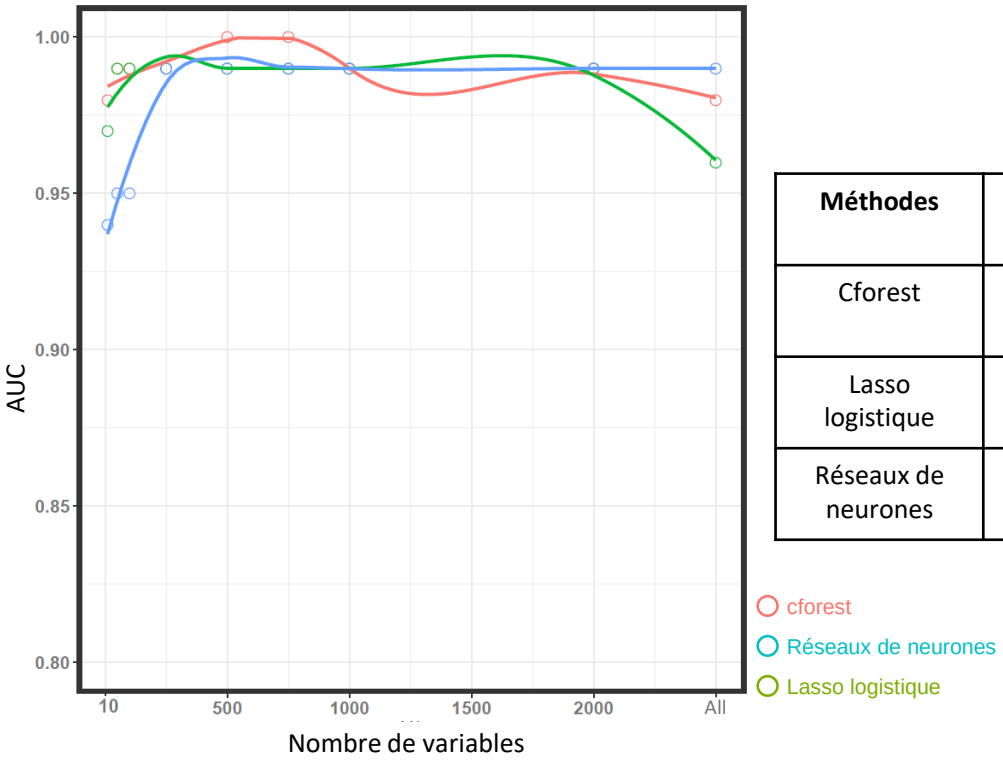
Sélection des variables

Une fois les hyper-paramètres définis, les modèles de prédiction ont pu être construits. Dans le travail sur la cohorte Holstein, l'intégration des différents types de données dans le but d'améliorer la prédiction des modèles n'avait que peu d'intérêt au vu de la bonne performance des modèles obtenus en n'utilisant que la méthylation de l'ADN comme source de variables. Néanmoins, ce travail a été réalisé dans le but d'identifier le nombre de variables minimal permettant de maximiser la prédiction. Pour cela, un premier modèle a été construit pour chaque méthode, avec toutes les variables disponibles. Cela a permis de classer les variables les plus contributrices, afin d'en extraire un échantillon réduit dans le but de générer de nouveaux modèles. Les résultats sont présentés dans la Figure 42. Comme attendu, on remarque que quelle que soit la méthode, les performances des modèles sont très bonnes, et stables quel que soit le nombre de variables incluses dans les modèles. Le nombre minimal de variables permettant de maximiser la prédiction est présenté pour chaque méthode en Figure 42.

Nature des variables sélectionnées par chaque méthode

Les variables sélectionnées par les méthodes décrites dans la partie précédente ont dans un premier temps été analysées afin de déterminer si elles étaient communes, et les résultats sont présentés en Figure 43 par un diagramme de Venn. Comme en race Montbéliarde, on observe une faible intersection entre les variables identifiées par les différentes méthodes, soulignant les spécificités de chaque méthode.

Nous nous sommes ensuite intéressés aux types de variables sélectionnées par chacune des méthodes. Encore une fois, on observe qu'aucune variable appartenant aux paramètres spermatiques n'a été sélectionnée, soulignant l'indépendance de ce type de variables vis-à-vis de la fertilité des animaux de la cohorte. Les autres types de variables ont tous été sélectionnés en proportions différentes selon les méthodes. Les réseaux de neurones et le cforest sélectionnent en majorité des SNP, néanmoins la



Méthodes	AUC	Nombre de variables
Cforest	1	500
Lasso logistique	0.99	50
Réseaux de neurones	0.99	250

Figure 42 : Performances des différents modèles en fonction du nombre de variables incluses. A gauche, l'AUC est représentée en y, et le nombre de variables les plus contributives ayant permis d'obtenir cette prédiction en x. Chaque courbe correspond à une méthode différente et a été obtenue en utilisant la fonction geom_smooth du package ggplot2. A droite est représentée une table qui référence pour chaque méthode l'AUC maximale et le nombre minimal de variables ayant permis d'obtenir cette AUC.

proportion des SNP est encore plus importante pour les réseaux de neurones. Le logistique lasso sélectionne principalement des variables appartenant aux CpG et aux sncRNA.

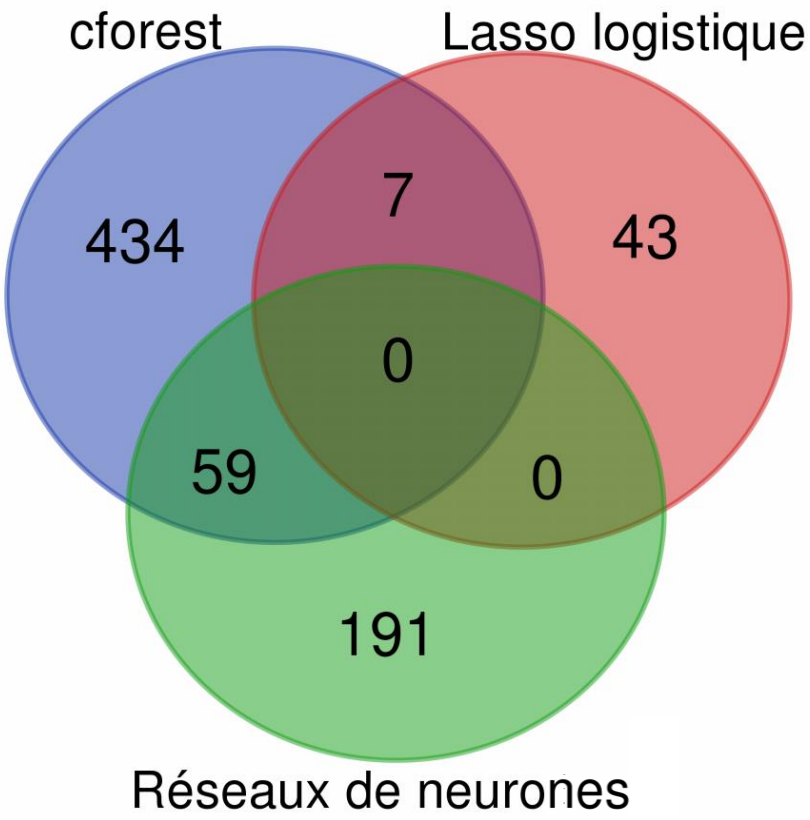
Analyse biologique des variables

La dernière partie de ce travail a consisté à analyser les différentes variables sélectionnées dans le but de déterminer si les gènes correspondants avaient été décrits en lien avec la fertilité. La démarche effectuée pour cette analyse est la même que dans la race Montbéliarde (Figure 7A de l'article 2). Les trois méthodes ont été analysées ensemble et cela représente 702 variables uniques. Les gènes contenant des SNP et des CpG sélectionnés ont été analysés ensemble, car ils sont directement ciblés par les variables et les gènes cibles des miRNA ont été analysés de leur côté (cibles potentielles).

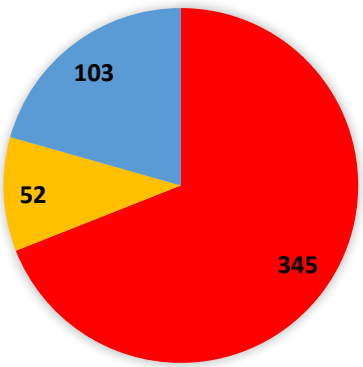
L'union des SNP et des CpG représente 639 variables uniques, correspondant à 341 gènes référencés sous Ensemble. Dans le but d'analyser leur fonction, une analyse d'enrichissement fonctionnel a été réalisée en utilisant DAVID, contre l'union des gènes couverts par l'ensemble des SNP et des CpG. A la suite de cette analyse d'enrichissement, 6 groupes de termes différents ont été mis en évidence et sont représentés dans la Figure 44A. Les groupes 2,4,5 et 6 sont respectivement associés à l'activité de l'ARN polymérase 2, la transcription, la localisation membranaire et l'adhésion cellulaire, fonctions biologiques indispensables dans un grand nombre d'organismes et de types cellulaires différents. Les groupes de termes 1 et 2 sont constitués de gènes ayant respectivement des domaines protéiques de type « Pleckstrin homology domain » et « PAS domain ». Ces deux domaines sont principalement impliqués dans la localisation intracellulaire des protéines et comme senseurs. Les SNP et les CpG sélectionnés correspondent donc à des gènes ayant des fonctions cellulaires essentielles dans de nombreux processus biologiques. De plus, parmi les 341 gènes pris en compte par DAVID, 92 ont des fonctions dans le développement.

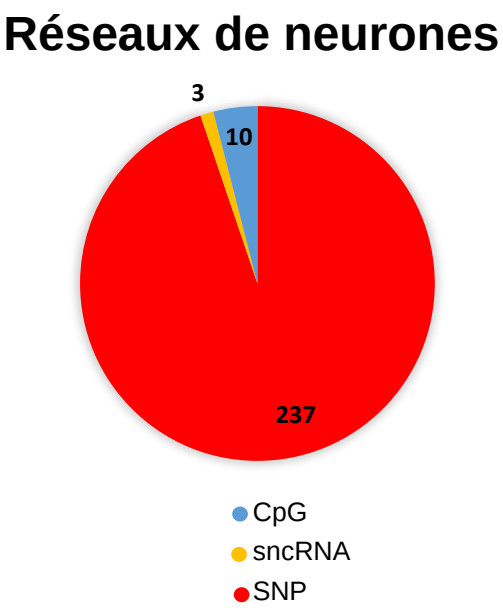
Enfin, nous nous sommes intéressés aux miRNA. Au cours de la sélection de variables 3 miRNA ont été identifiés par les modèles : miR-192, miR-146 et miR-16. Comme dans la race Montbéliarde, nous avons identifié les gènes cibles de ces derniers en utilisant le logiciel Targetscan afin de procéder à un

A



B





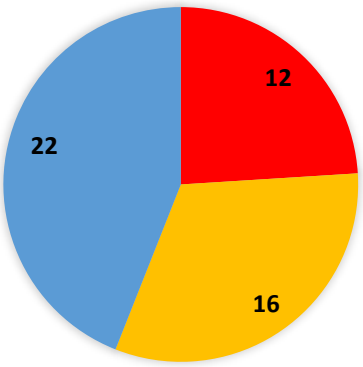


Figure 43 : Analyse des variables sélectionnées par les trois méthodes. A : Diagramme de Venn représentant les éléments communs ou spécifiques de chaque méthode. B : Pour chaque méthode les types de données sélectionnées sont représentés à l'aide d'un diagramme.

enrichissement fonctionnel. Seul le miR-192 et le miR-146 avaient des gènes cibles référencés sous Targetscan et 489 gènes cibles potentiels de ces miRNA ont été identifiés. Une analyse d'enrichissement fonctionnel de ces gènes a été réalisée en utilisant DAVID. En s'intéressant à la charte des annotations de la catégorie « processus biologiques », on remarque différents termes significativement enrichis directement en lien avec le phénotype d'intérêt. En effet, on observe les termes : « developmental process », « growth », « reproductive process » et « reproduction » qui nous indiquent qu'une partie des gènes cibles potentiels sont en lien avec la reproduction et le développement et donc avec la fertilité (Figure 44B).

Discussion

En comparant les résultats en race Montbéliarde et en race Holstein lors de cette intégration de données, on peut remarquer plusieurs points communs.

En effet, les résultats obtenus par AFM dressent des conclusions assez similaires. Tout d'abord, sur le jeu de données présélectionnées, on remarque que la première composante de l'AFM est liée à la fertilité, témoignant du fait qu'une partie de la variance du jeu de données est liée à cette dernière. On remarque que ce sont les CpG, les sncRNA et les SNP qui sont liés à cette dernière et pas les paramètres spermatiques, témoignant de leur indépendance vis-à-vis de la fertilité des animaux. En sélectionnant les variables quantitatives liées à la première composante on remarque une structure d'indépendance entre elles. Comme discuté dans l'article en race Montbéliarde, cela peut être due au profil transcriptomique particulier des spermatozoïdes, qui subit des modifications lors du transit épididymaire. Ces modifications ne sont pas liées à des mécanismes épigénétiques de régulation de l'expression génique dans le spermatozoïde, mais sont apportées par des épididymosomes. Il n'est donc pas surprenant que l'expression des sncRNA soient indépendante de la méthylation de l'ADN. Néanmoins, de par leur nature qualitative les SNP n'ont pas été explorés dans cette analyse, alors qu'il pourrait être intéressant d'observer s'ils sont liés à l'expression des sncRNA ou à un niveau de méthylation en particulier. Les résultats préliminaires d'inférence de réseaux réalisés dans la race

A

	Base de données	Terme	Gènes	Benjamini
Cluster 1 (EASE = 2.66)	INTERPRO	Pleckstrin homology-like domain	17	6.30E-01
	UP_SEQ_FEATURE	DOMAIN:PH	13	2.20E-01
	INTERPRO	Pleckstrin homology domain	13	6.30E-01
	SMART	PH	13	8.00E-01
Cluster 2 (EASE 2.65)	GOTERM_MF_DIRECT	RNA pylumerase II core promoter proximal region sequence-specific DNA binding	34	5.60E-02
	GOTERM_MF_DIRECT	regulation of transcription from RNA polymerase II promoter	38	7.10E+01
	GOTERM_MF_DIRECT	transcriptional activator activity, RNA polymerase II transcription regulatory region sequence-specific binding	17	1.10E-01
	GOTERM_MF_DIRECT	RNA polymerase II transcription factor activity, sequence-specific DNA binding	27	6.10E-01
	GOTERM_CC_DIRECT	nucleus	74	1.00E+00
Cluster 3 (EASE = 1.56)	INTERPRO	PAS domain	4	1.00E+00
	INTERPRO	PAC motif	3	1.00E+00
	SMART	PAC motif	3	1.00E+00
Cluster 4 (EASE = 1.51)	UP_KW_BP	Transcription	25	2.30E-01
	UP_KW_BP	Transcription regulation	23	3.20E-01
	UP_KW_CC	Nucleus	40	1.00E-01
Cluster 5 (EASE = 1.43)	UP_KW_CC	Membrane	118	1.60E-01
	GOTERM_CC_DIRECT	Integral component of membrane	90	1.00E-01
	UP_KW_DOMAIN	Transmembrane	93	3.80E-01
	UP_SEQ_FEATURE	TRANSMEM:Helical	90	1.00E-01
	UP_KW_DOMAIN	Transmembrane helix	68	9.60E-01
Cluster 6 (EASE = 1.36)	GOTERM_BP_DIRECT	cell adhesion	12	1.00E-01
	UP_KW_CC	Cell junction	11	8.70E-01
	UP_KW_BP	Cell adhesion	7	1.00E-01

B

Terme	Gènes	BH
development process	207	2.70E-07
growth	47	6.10E-06
cellular component organisation or biogenesis	214	6.10E-06
biological regulation	345	1.20E-05
single-organism process	367	3.40E-04
localization	191	2.40E-03
multicellular organismal process	214	3.60E-03
signaling	189	3.60E-03
response to stimulus	249	5.60E-03
cellular process	413	7.30E-03
locomotion	62	5.30E-02
behavior	22	6.50E-02
reproductive process	46	6.50E-02
reproduction	46	6.50E-02
biological adhesion	45	8.20E-02

Figure 44 : Analyse d'enrichissement fonctionnel. A : Groupes de termes enrichis parmi les gènes contenant les SNP et les CpG sélectionnés, par rapport à l'union des gènes couverts par l'ensemble des SNP et des CpG (« enrichment score » > 1.3). B : Charte d'annotation des gènes prédits comme cibles potentielles des miRNA sélectionnés.

Montbéliarde montrent que les interactions les plus fortes dans le jeu de données se font principalement au sein d'une couche de données –omiques et pas entre couches. Ces résultats, qui restent à confirmer sur plus de variables, suggèrent qu'il existe peu de lien entre les SNP et les autres types de données.

Nous avons ensuite sélectionné les variables qui permettaient de maximiser la prédiction au sein de chaque méthode. Dans cette sélection, on observe certaines divergences entre les races, tout d'abord en ce qui concerne le nombre de variables sélectionnées. En effet, le cforest sélectionne 100 variables en Montbéliarde mais 500 en Holstein ; le Logistic Lasso 250 en race Montbéliarde et seulement 50 en race Holstein. Une des raisons potentielles à l'origine de cette différence pourrait être la difficulté à fixer des seuils de sélection en race Holstein. En effet, en race Montbéliarde pour la plupart des méthodes, le nombre optimal de variables se traduit par un pic de performance des modèles de prédictions, ce qui facilite l'identification d'une valeur maximisant la prédiction. Néanmoins en race Holstein, les performances varient très peu : sur toutes les combinaisons testées, les pires performances ont été obtenues par les réseaux de neurones avec 10 variables, et l'AUC était de 0.94, ce qui reste quand même très bon. Dans notre cas, nous avons choisi un indicateur objectif : le nombre de variables minimal maximisant l'AUC ; néanmoins au vu des prédictions, il pourrait être intéressant de modifier ce critère de sélection des variables les plus pertinentes. Par exemple, le même nombre de variables qu'en Montbéliarde pourrait être proposé, dans le but de savoir si à nombre égal, la nature des variables pourrait varier selon la race.

Une autre différence entre les deux races, est le type de variables sélectionnées par les méthodes, en particulier les SNP. En race Holstein, les réseaux de neurones sélectionnent presque uniquement les SNP, qui sont majoritaires dans la sélection des cforest et représentent 1/3 des variables sélectionnées par les Lasso logistiques. De manière intéressante, le Lasso logistique évalue la contribution des variables de manière individuelle. Le fait que les SNP soient en proportion plus importants en race Holstein qu'en race Montbéliarde signifie qu'à l'échelle individuelle, ils discriminent mieux les taureaux

Holstein en fonction de leur fertilité. Cet aspect pourrait également expliquer pourquoi ils sont préférentiellement sélectionnés par les autres modèles dans la race Holstein comparé à la race Montbéliarde. Néanmoins, nous n'avons pour l'instant pas d'explication sur pourquoi les SNP sont de meilleurs prédicteurs en race Holstein qu'en race Montbéliarde.

Enfin, les analyses biologiques ont également pu mettre en évidence des divergences. Tout d'abord il n'y a aucun miRNA en commun entre ceux identifiés en race Holstein et ceux identifiés en race. Malgré cette absence d'intersection, on remarque que les gènes cibles de ces miRNA sont impliqués dans le développement.

Les groupes de termes enrichis parmi les gènes contenant des CpG et des SNP sont également différents entre les deux races. En race Montbéliarde, les groupes de termes sont associés à des gènes à Homeobox, impliqués dans le transport ionique ou présentant un domaine Fibronectin Type III. Alors qu'en race Holstein les groupes de termes enrichis concernent principalement la transcription, la localisation membranaire, l'adhésion cellulaire et la présence de domaine Pleckstrin ou PAS. Parmi les « Gene ID » contenant des CpG et SNP sélectionnés dans les deux races, seuls 20 sont en commun. Bien que cette intersection soit intéressante, car retrouvée dans les deux races, aucun terme GO dans la catégorie des « processus cellulaires » n'est enrichi et aucun enrichissement fonctionnel n'a pu être identifié pami cette liste. Cela ne permet donc pas de les lier au phénotype étudié. Ces résultats sont en accord avec ceux qui avaient été obtenus lors de l'analyse du méthylome spermatique (article 1, Costes *et al,.* 2022). Bien que certaines caractéristiques soient communes entre races, les gènes et les fonctions biologiques associés à ces gènes sont en général race-spécifique.

Discussion générale

Au cours de ce travail de thèse, nous nous sommes principalement intéressés à des différences de méthylome spermatique entre des taureaux fertiles et subfertiles de races Montbéliarde et Holstein. En réalisant des analyses différentielles de méthylation nous avons pu mettre en évidence qu'un grand nombre de DMC était localisé sur des polymorphismes de séquence. Cela a motivé la mise en place d'une stratégie adéquate pour prendre en compte ces polymorphismes de séquence, nous conduisant à supprimer les CpG potentiellement polymorphes. Après l'application de ce filtre et la réalisation de nouvelles analyses différentielles, nous avons identifié plusieurs centaines de DMC au sein de chaque race. Dans les deux races, nous avons pu mettre en évidence un déséquilibre de représentation entre les DMC et le background au niveau de certaines régions répétées. De plus, une partie de ces DMC se trouve dans des gènes ayant des fonctions décrites dans la physiologie du spermatozoïde et le développement embryonnaire et foetal. Enfin à partir des DMC sans données manquantes, il a été possible de construire des modèles de prédiction ayant une précision de 72% (Montbéliarde) et 94% (Holstein). En plus de ces éléments nous avons pu observer que le méthylome spermatique était relativement stable entre 13 et 19 mois ainsi que lors d'une déviation rapide de fertilité au cours de la carrière d'un animal.

Au cours du projet SeQuaMol, d'autres données ont été analysées, et parmi elles les sncRNA, certains paramètres spermatiques ainsi que les génotypes des animaux. Nous avons fait l'hypothèse que l'intégration de ces informations avec la méthylation de l'ADN pourrait permettre de mieux comprendre certains aspects moléculaires liés à la subfertilité des animaux, et dans le cas de la race Montbéliarde, d'augmenter la capacité de prédiction des modèles. Dans un premier temps, nous n'avons pas mis en évidence d'interactions importantes entre tables de données, y compris pour des variables très liées à la fertilité. Néanmoins, ces résultats doivent être approfondis par d'autres méthodologies comme l'inférence de réseaux. Nous avons également pu montrer dans les deux races que les paramètres spermatiques n'étaient pas de bons indicateurs des différences de fertilités entre les animaux de la cohorte. Enfin, à partir d'une combinaison de plusieurs types de variables, il est

possible d'augmenter la capacité de prédiction des modèles en race Montbéliarde jusqu'à atteindre une AUC de 0, 99.

Dans leur ensemble, ces résultats montrent qu'à partir d'informations moléculaires extraites du spermatozoïde et d'information génétique, il a été possible de prédire avec une bonne précision le statut de fertilité des animaux de la cohorte. Au cours de ce dernier chapitre j'aimerais mettre en perspectives ces résultats en abordant trois axes différents :

(i) J'aimerais dans un premier temps revenir sur certaines limites de ce travail, en particulier vis-à-vis des cohortes analysées, des données de méthylation de l'ADN et de leur interprétation biologique.

(ii) J'aimerais ensuite partager quelques considérations autour de l'application de mes résultats dans un objectif de prédiction de la fertilité sur le terrain.

(iii) Enfin, j'aimerais me pencher sur les perspectives d'utilisation et d'évolution d'un outil de routine basé sur des données épigénétiques spermatiques.

Cohortes analysées

Les cohortes ayant permis de construire les modèles de prédiction sont les cohortes « fertilité » en races Holstein et Montbéliarde. Ces cohortes peuvent être considérées comme de grande dimension ou de petite dimension en fonction du champ thématique considéré. Du point de vue de la physiologie, ces cohortes sont de grande dimension. Elles comprennent en effet 100 taureaux en race Montbéliarde et 38 en race Holstein. C'est à notre connaissance les plus grandes cohortes pour lesquelles des analyses pan-génomiques de la méthylation de l'ADN ont été menées dans ce domaine et pour l'espèce bovine, d'où leur caractère précieux. Ces cohortes permettent donc d'obtenir des résultats robustes, d'autant plus que les critères d'inclusion étaient assez stringent, avec le contrôle de l'âge à la collecte de l'éjaculat, le mélange de plusieurs éjaculats collectés dans une courte période de temps pour limiter les effets liés à une baisse de qualité ponctuelle, ainsi que le contrôle de la contamination somatique et des aléas techniques. Néanmoins, du point de vue des statistiques, les cohortes d'analyse

peuvent être considérées comme de faible dimension. Un faible effectif peut être source de biais dans la constitution de l'échantillon, qui ne serait pas représentatif des populations analysées. Il faut nuancer ce facteur, car les cohortes représentent une part importante de l'ensemble des classes d'âges analysées (environ 20% en Montbéliarde et 13% en Holstein). Il est donc difficile de les étendre davantage tout en conservant une fertilité contrastée. Il faut également tenir compte de la disponibilité du matériel biologique : les taureaux ont été sélectionnés sur la base de leur TNR 56, calculé à partir de plusieurs centaines d'IA. Ces taureaux, qui ont une carrière courte, ne sont bien souvent plus en service lors de leur recrutement dans une cohorte, et pour certains, toute la semence disponible a été écoulée ou jetée avec l'impossibilité d'en produire à nouveau. La seule façon d'étendre le dispositif est donc d'attendre que de nouveaux taureaux arrivent sur le marché, sachant qu'à l'heure actuelle seules quelques dizaines de nouveaux taureaux sont sélectionnés chaque année. Ainsi les résultats générés dans ce travail de thèse ne doivent pas être considérés comme immédiatement applicables par les entreprises de sélection afin d'identifier les taureaux subfertiles, mais plutôt comme des résultats encourageants démontrant que, sur la base d'informations génétiques et épigénétiques portées par le spermatozoïde, il est possible de prédire la fertilité bovine avec une très bonne précision, en particulier quand différentes sources de données sont prises en compte. Il reste néanmoins à réaliser des validations sur des cohortes de données plus larges en terme d'effectif, et indépendantes, afin de confirmer les résultats obtenus au cours de ce travail de thèse, en particulier dans la race Holstein. Ces cohortes indépendantes sont actuellement en cours de constitution.

Analyse des données de méthylation

Un des aspects qui a été central au cours de ce travail de thèse a été l'analyse des données de méthylation de l'ADN. Sur bien des aspects, ce type de données a été compliqué à analyser et cette partie de discussion a pour objectif de faire un retour d'expérience sur les éléments à la source de ces difficultés. Cette partie sera découpée en trois axes principaux : l'impact de la génétique sur les

données de méthylation, la fiabilité/reproductibilité de ces données et l'effet biologique de variations de méthylation.

Impact des génotypes

Au cours de la thèse, nous avons exploré l'impact de polymorphismes de séquence touchant directement des CpG, car c'est le facteur qui a généré quantitativement le plus de biais sur les données de méthylation. En analysant les résultats, nous avons pu observer que la valeur de méthylation associée à un CpG polymorphe reflétait la plupart du temps le nombre d'allèles portant un CpG méthylable. Nous avons également constaté que les DMC identifiées sans suppression au préalable des CpG polymorphes étaient en grande partie dans cette situation, montrant que ce biais affectait un grand nombre de CpG pertinents pour discriminer les taureaux fertiles et subfertiles. Ces DMC polymorphes étaient pertinents parce que des variations génétiques, et non épigénétiques, permettaient de discriminer les taureaux fertiles et subfertiles. Un certain nombre d'entre eux a d'ailleurs été ajouté à la nouvelle version de la puce de génotypage. Leur lien avec la fertilité pourra être étudié dans quelques années, quand suffisamment d'animaux auront été génotypés. L'identification majoritaire de CpG polymorphes par l'algorithme methylKit utilisé pour l'analyse différentielle montre que cet outil tend à sélectionner de la variabilité génétique plutôt qu'épigénétique. La publication de ces résultats (article 1, Costes *et al.,* 2022) devrait attirer l'attention de la communauté scientifique sur les biais de cet outil qui est très utilisé. A la lumière de ces résultats, nous avons choisi la solution qui nous semblait être le meilleur compromis entre suppression des sources de biais et conservation d'un maximum de données, en utilisant la base de données des 1000 Génomes Bovins référençant une grande partie des polymorphismes de séquences chez le bovin. Néanmoins, cette stratégie, bien qu'intéressante, n'est pas optimale. Cette solution est très drastique, puisque l'on supprime des CpG potentiellement polymorphes dans l'ensemble de la population bovine ; cependant, rien ne nous dit que les animaux analysés au cours de la thèse présentent effectivement des polymorphismes à ces sites. On pourrait tout à fait envisager que certains

polymorphismes avec des fréquences alléliques faibles ne soient pas représentés parmi nos individus. De plus certains allèles avec des fréquences rares auront peu d'impact sur les données de méthylation même s'ils sont présents dans la cohorte d'analyse. On pourrait donc envisager des sources d'amélioration des filtres utilisés dans ce travail pour cibler les polymorphismes impactant de manière importante les données de méthylation. Pour ces raisons, la méthode utilisée dans ce travail est justifiée mais pas optimale. La meilleure façon de se débarrasser intégralement de ce biais serait bien évidemment d'avoir accès à la séquence génomique de chaque animal. Dans le cas de la technique du RRBS, il est possible d'avoir accès à l'information de séquençage de la guanine d'un site CpG. Au moyen d'un traitement bioinformatique supplémentaire, il pourrait ainsi être envisagé de supprimer de l'analyse la moitié des CpG polymorphes dans les cohortes analysées (Liu *et al.*, 2012). Le cas des polymorphismes affectant la cytosine d'un site CpG est plus compliqué. En effet dans la majorité des cas, l'allèle alternatif d'une cytosine localisée dans un site CpG est une thymine. Ainsi lors du séquençage il n'est pas possible de pouvoir faire la différence entre une thymine « d'origine » et une thymine issue de la transformation d'une cytosine non méthylée lors de la conversion en bisulfite. La technologie du RRBS n'est donc pas idéale pour intégrer une information de génotypes dans le cadre de la correction des données de méthylation.

Il existe des technologies d'analyse de la méthylation non basées sur une conversion chimique ou enzymatique entrainant une confusion entre signal de méthylation et polymorphisme C>T. C'est le cas par exemple les technologies de type Oxford Nanopore Technology (Branton *et al.*, 2008). Ces dernières peuvent faire une distinction entre thymine, cytosine, methylcytosine et même hydroxyméthylcytosine. Utiliser cette technologie permettrait donc d'avoir accès à la méthylation de l'ADN sans confusion avec les polymorphismes de séquences.

Néanmoins, la présence d'un CpG polymorphe n'est pas le seul élément d'ordre génétique interférant avec les données de méthylation de l'ADN. Il a en effet été montré dans de nombreuses études que le profil de méthylation de l'ADN était un caractère héritable chez l'Homme, avec une héritabilité variant

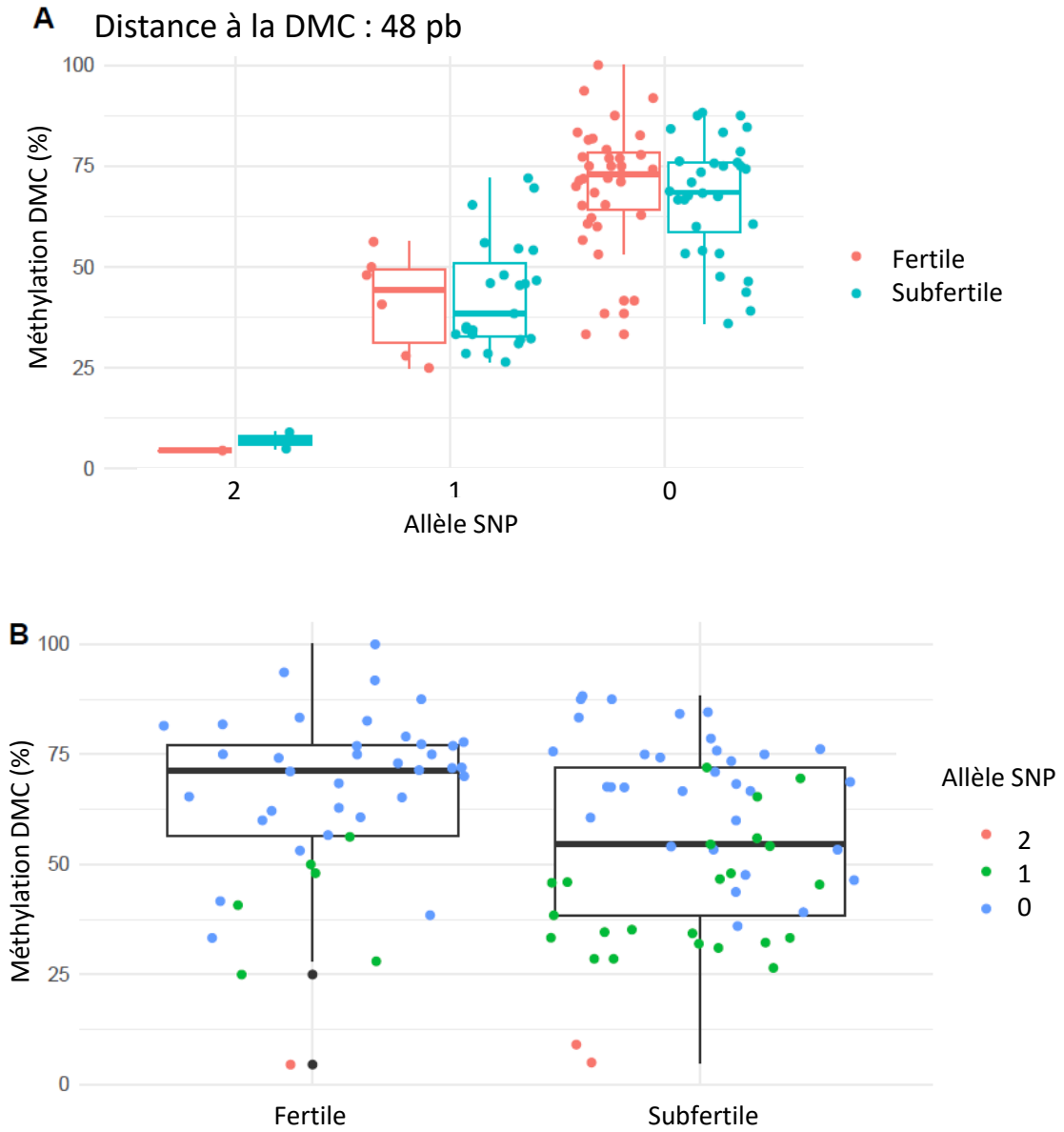
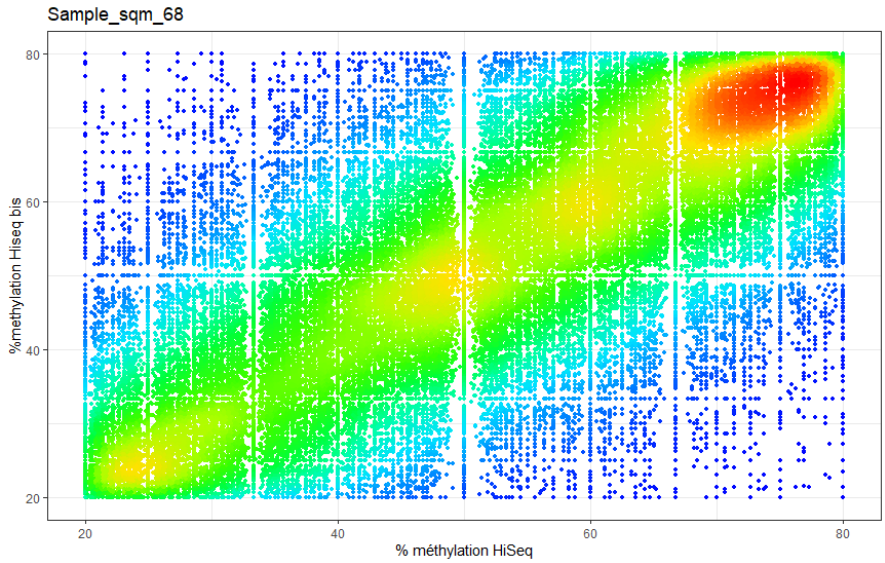
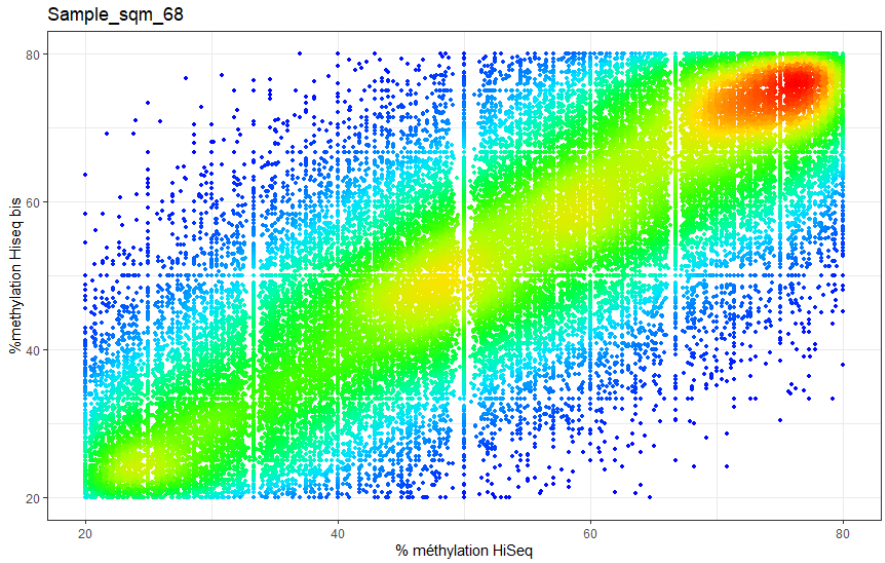


Figure 45 : Analyse de l'impact d'un polymorphisme d'un CpG sur le niveau de méthylation d'une DMC. A : Boxplot représentant le niveau de méthylation d'une DMC (en y) en fonction de l'allèle d'un CpG polymorphe distant (en x). Deux boxplots ont été tracés par modalité allélique en fonction de la fertilité. Dans cette figure on observe bien un impact du polymorphisme du CpG distant sur le niveau de méthylation. B : Représentation en boxplot du même phénomène, le niveau de méthylation de la DMC est représentée sur l'axe y mais la fertilité est présente sur l'axe x. Les points de données correspondant à chaque animale sont coloriés en fonction de l'allèle du CpG distant. On observe bien, que ce soit chez les animaux fertiles et subfertiles, une corrélation importante entre le niveau de méthylation et l'allèle du CpG distant.

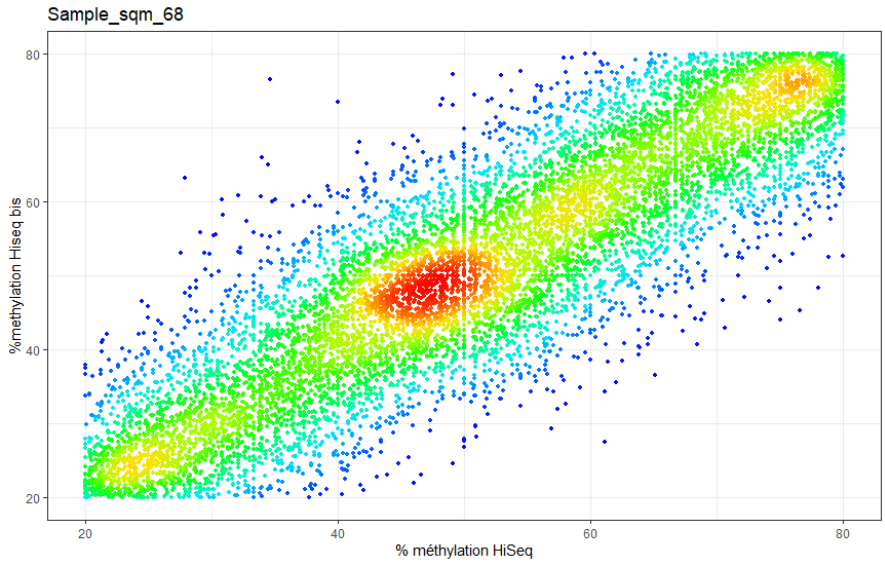
entre 16 et 80% (Gertz *et al.*, 2011; Bell *et al.*, 2012; Grundberg *et al.*, 2013; Taudt *et al.,* 2016; Lappalainen and Greally, 2017). Bien que ces résultats fluctuent beaucoup entre les études, du fait des différences de cohortes et de panels de CpG ciblés, il y a un consensus sur le fait qu'il existe un contrôle génétique du méthylome. Ces régulations peuvent se faire de manière *cis* ou *trans* (c'est-à-dire à plusieurs Mb de distance ou sur 2 chromosomes différents) (Do *et al.*, 2017). Ainsi, la présence d'un polymorphisme distant d'un CpG pourrait contrôler son niveau de méthylation en fonction du génotype auquel il est associé, et donc engendrer une source de biais qui n'a pas été prise en compte par notre stratégie de filtrage des CpG polymorphes. Nous avons voulu examiner si certaines de nos DMC étaient impactées de manière *cis* par des SNP. Pour cela, nous avons fait une recherche exploratoire dans la race Montbéliarde. Nous avons regardé si le CpG polymorphe le plus proche de la DMC analysée favorisait un certain niveau de méthylation en fonction de son génotype. Pour certaines positions, on observe effectivement un lien entre génotype et méthylation de l'ADN, démontrant bien l'existence de ce biais. Un exemple de CpG non polymorphe se trouvant dans cette situation est montré Figure 45 : l'impact du polymorphisme en *cis* sur la distribution de méthylation est moindre par rapport à un CpG polymorphe. En effet, la distribution du niveau de méthylation n'est pas discrète et il existe une grande variabilité intra-génotype, ce qui suggère une interaction entre facteurs génétiques et non génétiques dans le contrôle du niveau de méthylation de ce CpG. Il est donc difficile de déterminer si l'impact des DMC identifiées sur la fertilité a pour cause des facteurs génétiques indirects, la méthylation de l'ADN non contrôlée par des facteurs génétiques, ou une interaction entre ces deux entités. Pour corriger les données de méthylation de ces facteurs génétiques indirects, il faudrait mettre en œuvre une stratégie plus compliquée que le filtre des CpG polymorphes que nous avons effectué. Afin de le faire en toute rigueur, il faudrait disposer d'une cartographie précise de toutes les séquences en *cis* ou en *trans* contrôlant les niveaux de méthylation à un CpG donné (QTL de méthylation), ce qui n'existe actuellement pas chez le bovin et nécessite des cohortes de très grande taille (500 individus).



N>10
74 146 CpG
R^2 = 0,48



N>20
41 032 CpG
R^2 = 0.61



N>50
9 508 CpG
R^2 = 0.78

Figure 46 : Variation du niveau de méthylation sur un réplicat technique. Un graphique représentant sur l'axe x et l'axe y le niveau de méthylation de CpG d'un réplicat technique. Chaque point correspond à la valeur de méthylation pour un CpG donné. Le nombre de point représenté était grand, ainsi ils sont coloriés en fonction de la densité de point a proximité. Les graphiques représentent les niveaux de méthylation intermédiaire (entre 20 et 80% de méthylation). Les graphiques ont été obtenues en sélectionnant un niveau de couverture minimale différent (respectivement 10, 20 et 50). On observe, qu'a un seuil de couverture de 10, la corrélation est assez faible et qu'elle augmente en fonction de la couverture des CpG.

A la lumière de ces deux observations, on peut donc conclure que la génétique impacte à bien des égards les données de méthylation de l'ADN, et que la stratégie mise en œuvre dans la thèse n'a permis qu'une correction partielle de ces effets génétiques.

Fiabilité des résultats de RRBS

Un autre aspect limitant dans l'analyse de la méthylation de l'ADN, qui est plus facile à évaluer que l'impact de la génétique, est la fiabilité des résultats de méthylation de l'ADN. Un des éléments techniques ayant un impact fort sur les données de méthylation en RRBS est la profondeur de séquençage des CpG. Cet élément n'est pas le seul garant de la qualité des données que l'on analyse, on pourrait également penser au taux de conversion au bisulfite des banques analysées, ainsi qu'à l'homogénéité de la sélection de taille (et donc de couverture génomique) entre échantillons.

La profondeur a une importance majeure pour la précision et la fiabilité de la mesure effectuée. En effet, la précision de la valeur de méthylation en dépend : avec 10 séquences couvrant un CpG, seules 11 valeurs de méthylation sont possibles (0, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%) ; la précision sera donc de 10% alors qu'avec 20 séquences elle sera de 5%. De plus, la profondeur est également importante pour estimer la fiabilité et la répétabilité d'une valeur de méthylation. En effet, avec une profondeur de 10 séquences, la probabilité de tirer un échantillon de molécules non représentatif de la proportion réelle de molécules portant un C ou un T dans la banque est naturellement plus élevée qu'avec une profondeur de 20 séquences. La conséquence de ces deux éléments fait que les données de méthylation peuvent être hautement variables en fonction des paramètres calibrés pour l'analyse bioinformatique. Pour illustrer cela, un réplicat technique de séquençage a été réalisé sur une même banque et est présenté en Figure 46. A un seuil de profondeur de 10 séquences, on remarque que l'écart de mesure entre les deux réplicats est très important pour les valeurs de méthylation intermédiaires (entre 20% et 80% de méthylation), comme le montre la faible valeur du coefficient de corrélation, et diminue à mesure que l'on augmente la profondeur. Ces

observations sur nos données de RRBS sont confirmées par une étude sur données simulées (Seiler Vellame *et al.*, 2021).

Ces résultats peuvent apporter un questionnement sur les paramètres bioinformatiques qui ont été appliqués au cours de mon travail de thèse puisque nous nous sommes placés à un seuil de profondeur de seulement 10 séquences. L'impact de ces filtres sur la précision de la valeur de méthylation des DMC doit être relativement négligeable puisque les DMC sont en général couvertes bien au-delà de ce seuil (Montbéliard : 24 et 36 séquences en moyenne pour les DMC respectivement avec et sans données manquantes ; Holstein : 38 et 43 séquences en moyenne pour les DMC respectivement avec et sans données manquantes). Cela peut être dû au fait que pour qu'un CpG soit éligible à l'analyse et puisse être intégré au background, ce seuil doit s'appliquer à au moins la moitié des animaux d'un groupe de fertilité (soit 22 en Montbéliarde et 9 en Holstein), privilégiant les régions génomiques bien couvertes. De plus, dans l'article 1, on observe une corrélation importante entre les niveaux de méthylation obtenus par RRBS et pyroséquençage (technique qui ne repose pas sur la profondeur de séquençage pour estimer le niveau de méthylation). Ces résultats suggèrent qu'au moins dans la race Montbéliarde, le niveau de méthylation des DMC est mesuré de façon précise.

Néanmoins, un séquençage plus profond se traduit par un coût plus élevé, qui sera souvent compensé en diminuant le nombre d'échantillons analysés. La stratégie à mettre en place dépend donc de la question biologique posée. Un séquençage peu profond et un nombre d'échantillons limité peuvent être suffisants pour mettre en évidence des différences de méthylation très importantes, comme existant par exemple entre deux types cellulaires Nous ne sommes clairement pas dans ce cas de figure, avec des différences moyennes de 14% (Montbéliarde) et de 15% (Holstein) au niveau des DMC identifiées. Ces deux paramètres (profondeur de séquençage et taille de l'échantillon) conditionnent tous deux la puissance statistique de l'étude (Seiler Vellame *et al.*, 2021). Sur des données simulées, il semblerait qu'une augmentation de la profondeur de séquençage n'ait que des effets limités sur la puissance nécessaire à la détection de différences de méthylation de 10% (qui est le seuil que nous

avons utilisé pour la détection des DMC). La taille de l'échantillon a un effet de levier bien plus important : pour un échantillon de 100 individus avec un effectif légèrement déséquilibré dans les 2 groupes (60/40, ce qui est similaire à la cohorte fertilité en race Montbéliarde), des différences de méthylation de 10% peuvent être détectées avec une puissance de 80%. Cette puissance n'atteint pas 25% pour un échantillon de 50 individus, ce qui est plus proche des conditions de la cohorte Holstein. Il est donc possible que le nombre de DMC soit sous-estimé dans cette race.

Impact biologique de variations de la méthylation de l'ADN

Enfin, le dernier aspect que j'aimerais aborder dans le cadre de l'analyse des données de méthylation de l'ADN, est leur interprétation biologique. Cet aspect n'est pas obligatoire lorsque l'on cherche par exemple à utiliser la méthylation de l'ADN uniquement comme un support permettant de réaliser une prédiction (biomarqueurs prédictifs). Il est néanmoins indispensable quand on essaye d'extraire du sens de ces données pour la compréhension du phénotype d'intérêt.

Dans un premier temps, je vais aborder l'analyse dans le cadre des régions géniques. Quand un CpG se trouve dans un gène ou à proximité, nous annotons ce CpG à ce gène, qui est ensuite exploité pour proposer des hypothèses sur les conséquences fonctionnelles du différentiel de méthylation. Pour que cette démarche soit valide, il faut que les conditions suivantes soient vérifiées:

(i) Que le CpG différentiellement méthylé ait réellement une action biologique sur la régulation de l'expression des gènes.

(ii) Que les CpG portant une méthylation différentielle au niveau d'un gène aient bien une action biologique sur ce gène. En effet, dans les représentations graphiques, l'ADN est souvent représenté dans un espace à une dimension. Cela est physiologiquement faux, la chromatine étant structurée de manière tridimensionnelle dans une cellule. Ainsi, un CpG différentiellement méthylé sur un gène pourrait en réalité avoir une action biologique à distance sur le chromosome, voir sur un autre chromosome, plutôt que sur le gène à proximité (Huan *et al.*, 2019).

(iii) Enfin il s'agit de comprendre la nature biologique du signal. Les CpG méthylés présents dans un promoteur de gène sont souvent associés à de la répression transcriptionnelle. Pour ce qui est des autres éléments géniques, l'effet d'une hyper ou hypométhylation est beaucoup moins général et moins bien documenté (Deaton and Bird, 2011; Lee *et al.*, 2017).

Ces trois points ci-dessus montrent qu'il est compliqué de réellement comprendre l'impact biologique d'une méthylation différentielle. Les hypothèses à la base de l'interprétation biologique de différences de méthylation sont donc fortes et pas toujours vérifiables. Dans la plupart des études portant sur la méthylation de l'ADN, y compris la nôtre, ces hypothèses sont posées, permettant de grandement simplifier le problème. Néanmoins, au vu du grand nombre de DMC identifiés dans les analyses différentielles, on peut raisonnablement penser que pour une partie des DMC au moins une de ces trois hypothèses n'est pas respectée, biaisant les déductions biologiques que l'on fait à partir des données de méthylation. Il est donc important d'être prudent dans les conclusions tirées. Une manière de pouvoir vérifier certaines des conséquences biologiques d'un différentiel de méthylation de l'ADN serait d'analyser d'autres entités biologiques, comme par exemple le transcriptome dont l'impact fonctionnel est plus clair. En réalisant une intégration de ces deux types de données (par inférence par exemple), il serait possible de lier la présence d'une méthylation différentielle, à une diminution ou augmentation de la transcription, voire à la présence de transcrits alternatifs dans un cadre général. Néanmoins, dans ce travail nous nous sommes intéressés au spermatozoïde qui est une cellule transcriptionnellement inactive. De plus, nous avons pu mettre en évidence qu'un grand nombre des DMC identifiées étaient associées à des gènes importants pour le développement. La méthylation de l'ADN est essentielle au développement embryonnaire (Li *et al.,* 1992; Okano *et al.*, 1999; Beaujean *et al.*, 2004). Ainsi pour réaliser des explorations fonctionnelles des DMC mises en évidence dans ce travail, il pourrait être intéressant de travailler sur l'embryon. Dans l'unité, des travaux en cours utilisent la semence des taureaux ayant une méthylation contrastée au niveau des DMC identifiées dans ce travail pour réaliser des fécondations *in vitro*. Le but de ce travail sera de suivre le début du

développement afin d'observer s'il y a des différences au niveau du taux de formation des blastocystes, et également de réaliser des profils transcriptomiques et d'analyser le méthylome des embryons. Néanmoins ces approches, bien qu'intéressantes, comportent un grand nombre de facteurs confondants (les semences utilisées ne diffèrent pas que sur le plan du méthylome), si bien qu'il ne sera pas possible de pouvoir dresser une relation de cause à effet entre différence de méthylation spermatique et développement embryonnaire. Il pourrait donc être intéressant de réaliser de l'édition de l'épigénome chez l'embryon, technique consistant à modifier le statut de méthylation de l'ADN à des locus précis dans le génome (Yamazaki *et al.*, 2020). Ainsi, en exploitant le polymorphisme de séquence entre les allèles paternel et maternel, il pourrait être possible de faire varier le niveau de méthylation de l'allèle paternel, en introduisant peu de facteurs confondants.

L'analyse des séquences répétées pose également des difficultés. Il convient de noter que ces difficultés ne sont pas spécifiques à l'analyse de la méthylation de l'ADN, mais sont une problématique partagée par toutes les analyses s'appuyant sur du séquençage en « short reads » suivi d'un alignement sur le génome. En effet, les séquences répétées sont représentées en plusieurs exemplaires sur tout le génome. Ainsi, lorsque des séquences correspondant à des séquences répétées sont alignées sur un génome de référence, il n'est pas rare qu'elles puissent être alignées à plusieurs endroits du génome. Au cours de l'analyse bioinformatique ces séquences multiples sont éliminées pour éviter toute ambiguïté. Cela a pour conséquence de négliger ce type de séquences. Pour prendre cela en considération, nous avons réalisé dans notre étude en race Montbéliarde, un alignement sur un génome composé d'un seul exemplaire de chaque région répétée (obtenu à partir de la base de données RepBase (Bao *et al.,* 2015)). Néanmoins, ce type d'alignement sous-tend que toutes les séquences d'un même type se comportent de la même façon dans tous les génomes, ce qui n'est pas toujours le cas. L'analyse de ces séquences est néanmoins particulièrement intéressante dans le cadre de l'analyse de la fertilité. En, effet, dans l'analyse des DMC nous avons pu constater un enrichissement en retrotransposons de types LINE. La régulation de ces éléments est particulièrement critique au moment du développement embryonnaire. En effet une expression de ces séquences est nécessaire

afin d'ouvrir la chromatine et rendre possible la transcription embryonnaire (Jachowicz *et al.*, 2017). Néanmoins, une expression trop importante de ces séquences entraine une instabilité génomique et est également néfaste pour le développement (Jachowicz *et al.*, 2017). Un des nombreux rôles de la méthylation de l'ADN est le contrôle de ces régions ; ainsi toute altération du méthylome pourrait potentiellement avoir des conséquences néfastes sur le développement (Kohlrausch *et al.*, 2022). Ainsi, une analyse plus précise des régions répétées est particulièrement pertinente dans le cadre de l'analyse de la méthylation de l'ADN et de la fertilité. Des technologies permettant de faire du séquençage en « long reads » permettraient en partie de résoudre ce problème, au détriment de la taille de la cohorte du fait de leur coût élevé.

Enfin il faut ajouter qu'en dehors des éléments géniques et répétés, le génome contient des éléments régulateurs dont l'activité est en partie régulée par la méthylation de l'ADN, tels que les « enhancers », les « silencers », les sites de fixation pour des facteurs de transcription ou les éléments insulateurs (Schübeler, 2015). On pourrait imaginer qu'une méthylation différentielle au niveau de tels loci dans la semence pourrait être transmise à l'embryon et interférer avec la mise en route du génome embryonnaire (Kremsky and Corces, 2020). Il est cependant difficile d'explorer la pertinence de cette hypothèse chez le bovin, où ces séquences régulatrices sont encore partiellement annotées.

Application pratique des résultats de la thèse sur le terrain

Prédiction de la fertilité au niveau du taureau ou de l'éjaculat

Dans la prédiction de la fertilité mâle, il y a deux enjeux : la prédiction au niveau taureau et la prédiction au niveau éjaculat. La prédiction au niveau taureau se caractérise par le fait d'obtenir une valeur témoignant de la fertilité de l'animal sur l'ensemble de sa carrière. L'avantage principal de cette mesure est qu'elle ne doit être réalisée qu'une seule fois dans la carrière de l'animal (le plus tôt possible), et donc elle présente peu de difficultés pratiques et engendre peu de coûts. Néanmoins l'inconvénient de cette démarche est que la fertilité d'un taureau peut évoluer au cours de sa carrière, et qu'à partir d'une seule mesure il est impossible de prédire cette dynamique. Ainsi il pourrait

également être intéressant d'avoir un prédicteur de la fertilité au niveau éjaculat afin d'être capable de prédire son évolution. Néanmoins cette évaluation à l'éjaculat comporte des inconvénients : il faut en effet mettre en place une logistique pour réaliser de manière efficace et fluide une grande quantité de tests, et il faut également que le coût de l'analyse reste raisonnable.

Ainsi la stratégie à adopter doit tenir compte de plusieurs facteurs dont l'objectif de prédiction, le coût de la technique expérimentale et l'évolution de la marque analysée au cours de la carrière de l'animal.

Nous avons pu observer en analysant la cohorte longitudinale « âge » que le méthylome spermatique évoluait peu entre 13 et 19 mois. Le biais que peut avoir l'âge des animaux sur le niveau de méthylation est assez faible et une information de méthylation acquise à 13 mois semble donc représentative du méthylome jusqu'à 19 mois à quelques CpG près. En ce qui concerne l'analyse de la cohorte « déviation », nous avons vu qu'un changement rapide de la fertilité se traduit par un nombre négligeable de DMC (25 DMC au maximum).

Dans l'ensemble, ces résultats montrent que le méthylome spermatique est une marque relativement stable chez le taureau de 13 à 19 mois et qu'elle n'est pas liée à des variations rapides de la fertilité. L'ensemble des résultats montre que la méthylation de l'ADN ne peut pas être utilisée pour un suivi de fertilité au niveau éjaculat. Néanmoins, de par sa stabilité elle pourrait être utilisée une fois en début de carrière, afin de réaliser une prédiction au niveau taureau.

Quels types de données -omiques choisir pour prédire la fertilité ?

Une question que l'on peut également se poser au vu des résultats obtenus à l'issue de ce travail de thèse, concerne les types de données à utiliser pour la prédiction de la fertilité mâle. Il manque quelques résultats essentiels afin de répondre entièrement à cette question. Néanmoins on peut faire émerger certaines directions, qui seront bien évidemment à confirmer. Dans cette discussion je vais principalement m'appuyer sur les résultats en race Montbéliarde car la cohorte principale était plus grande et que nous avions également à disposition une cohorte indépendante pour l'analyse de la

méthylation de l'ADN. Les résultats en race Holstein ne sont pas insatisfaisants, bien au contraire ; mais des données supplémentaires de validation sont nécessaires pour placer un bon niveau de confiance dans les résultats obtenus.

Tout d'abord, en se basant uniquement sur les données de méthylation de l'ADN, nous avons pu construire des modèles avec une précision de prédiction de 72% dans la cohorte principale et dans la cohorte indépendante. Ces résultats démontrent qu'il est possible d'avoir une prédiction non aléatoire en se basant sur les DMC identifiées, mais que pour certains animaux, le profil de méthylation de ces DMC ne permet pas de les classifier correctement. L'hypothèse émise pour expliquer ce phénomène repose sur le fait que la fertilité est un caractère multi-factoriel et donc analyser un seul de type de données ne garantit pas de pouvoir expliquer l'intégralité de ce phénotype. C'est pour cela que cette analyse s'est poursuivie par une analyse intégrative des données de méthylation de l'ADN, des sncRNA, des SNP et des paramètres spermatiques. Les résultats montrent que la qualité de prédiction obtenue par cette analyse intégrative est bien supérieure pour toutes les méthodes utilisées, et quasiment parfaite avec la méthodologie des réseaux de neurones. On remarque ainsi qu'en apportant d'autres sources d'information la qualité de prédiction augmente par rapport à l'utilisation de la méthylation seule. A la lumière de ces résultats on peut donc préférer une approche intégrative par rapport à l'analyse d'un seul type de données -omiques.

Néanmoins, pour une application sur le terrain, plus l'on analyse de types de variables, plus l'on doit mettre en place de techniques différentes ce qui engendre des difficultés organisationnelles et une augmentation des coûts financiers du test. La question à se poser est donc la valeur ajoutée d'un nouveau type de données par rapport au gain de précision de la prédiction et à l'investissement technique et financier à réaliser. Il est difficile de répondre complétement à cette question sans réaliser de prédiction sur chaque table de données individuellement ni deux à deux avec la même approche appliquée à chaque type de données, ce que nous n'avons pas complètement finalisé au cours de la

thèse. Il est indispensable aussi d'avoir une vision assez précise du coût de chaque test dans une perspective d'utilisation de routine.

Néanmoins, il faut prendre en compte le fait que les SNP sont un type de données accessible sans coût supplémentaire. En effet, les taureaux commerciaux utilisés en insémination artificielle ont été sélectionnés comme reproducteurs basé sur leur génotype. Nous avons pu voir que le génotype des animaux contenait des variables souvent contributrices des modèles prédictifs. Nous pourrions donc déjà inclure cette donnée au modèle. Le choix d'un type de données supplémentaire doit donc se faire entre méthylation de l'ADN et sncRNA. Nous avons pu observer dans cette thèse que la méthylation de l'ADN seule permettait de prédire correctement la fertilité ; de plus dans les analyses intégratives les variables de type CpG ont souvent été sélectionnées. Cette conclusion est également vraie pour les sncRNA qui ont été analysés de manière similaire dans le travail de thèse d'E. Sellem. Ainsi utiliser l'une ou l'autre de ces types de données en combinaison avec les génotypes permettrait potentiellement de bonnes performances de prédiction. L'analyse des cohortes longitudinales « déviation » et « âge » n'a pas été terminée pour les sncRNA, nous n'avons donc pas de recul sur leur potentiel d'utilisation au niveau taureau ou éjaculat. Cependant, il est tentant de spéculer que les sncRNA spermatiques, dont le contenu peut rapidement être modulé au cours du transit épididymaire, sont plus labiles que la méthylation de l'ADN et pourraient être utilisés pour prédire la fertilité au niveau éjaculat, une fois la prédiction au niveau taureau réalisée à l'aide de la méthylation de l'ADN. Un sncRNA détecté dans les spermatozoïdes matures a par exemple pu être associé à une subfertilité suivant une infection du tractus génital chez l'homme (Chu *et al.*, 2017). Selon le coût et la rapidité d'exécution du test à appliquer, des sncRNA pourraient être mesurés de manière systématique ou à intervalle régulier, ou pourraient être mesurés après une maladie, de manière à s'assurer que celle-ci n'a pas eu de répercussion négative sur la fertilité des éjaculats collectés.

Quelles perspectives pour la prédiction de la fertilité mâle à partir de données épigénétiques ?

A la question : « Faut-il mieux estimer la fertilité des taureaux ? », la réponse est évidemment oui. Nous avons pu voir dans l'introduction que la commercialisation de semence subfertile a des conséquences non seulement économiques, mais également sur l'organisation des systèmes d'élevage. Maitriser la fertilité d'un taureau permettrait donc d'éviter en partie ces problèmes. La réelle question est donc plutôt de savoir quel est le coût maximum que des acteurs de l'élevage peuvent investir afin de prédire la fertilité des animaux. En d'autres termes, deux éléments sont à prendre en compte : combien peut rapporter une meilleure prédiction de la fertilité des animaux et combien cela coûte. Pour le premier point, à ma connaissance, il n'existe pas de modélisation économique permettant de connaitre le gain économique d'un point de fertilité chez les taureaux. Il est donc difficile de peser correctement la balance bénéfice/coût. Néanmoins certaines entreprises ont mis au point des évaluations de la fertilité de taureaux basées sur les spectres mIR (spectrométrie moyen infrarouge) de la semence (information sur la technologie : https://www.evolution-xy.fr/fr/actualite/nouveau-decouvrez-fertimax). Bien que le coût de ce service reste confidentiel, l'acquisition de spectre mIR est une technologie peu coûteuse, en tout cas moins coûteuse que l'analyse de la méthylation de l'ADN par RRBS. De plus ces technologies offrent l'avantage d'avoir une prédiction de la fertilité au niveau de l'éjaculat et pas seulement au niveau du taureau. Elle peut ainsi permettre un suivi de la fertilité des animaux au cours de leur carrière.

A la lumière de ces éléments, on peut donc réellement se poser la question de l'intérêt de l'utilisation de données épigénétiques (plus chères et avec un suivi de la fertilité plutôt au niveau taureau, en tout cas pour la méthylation de l'ADN), par rapport à d'autres types de données. Certes, d'un point de vue fondamental, la compréhension de l'implication du méthylome spermatique dans la fertilité est d'un intérêt indiscutable, mais est-ce que son intérêt dans la mise en place d'un outil de diagnostic de la fertilité l'est également ? Si l'objectif est de développer une évaluation de la fertilité, alors l'analyse

des spectres miR peut être suffisante. Il serait d'ailleurs intéressant de vérifier l'efficacité des prédictions obtenues à l'aide de ces données sur nos cohortes. Il est cependant important de noter que l'information épigénétique, en plus de son utilisation pour prédire la fertilité, pourrait également avoir d'autres applications pratiques.

Héritage du méthylome spermatique à l'embryon

L'héritage partiel du méthylome spermatique à l'embryon ouvre des perspectives allant au-delà de la prédiction de la fertilité mâle. En effet, comme présenté dans l'introduction de ce manuscrit, à l'issue de la fécondation, le méthylome gamétique est rapidement déméthylé afin de permettre à l'embryon de retrouver un état totipotent. Néanmoins, une partie du méthylome spermatique résiste à cette déméthylation, dont des éléments transposables de type IAP (chez la souris), LINE et également des régions soumises à l'empreinte parentale. Ces régions ont fait l'objet de nombreuses études afin de comprendre leur implication dans le développement.

Néanmoins, ces régions ne sont pas les seules à être transmises à l'embryon. En effet, une partie du niveau de méthylation est sous contrôle génétique et donc par définition également héritable (Taudt, Colomé-Tatché and Johannes, 2016).

De plus, il a été montré que d'autres régions pouvaient résister à la reprogrammation épigénétique (Zheng *et al.*, 2021). Les auteurs de cette étude ont soumis des souris à un stress de contention sur une longue période et ont pu mettre en évidence un certain nombre de DMR dans leur méthylome spermatique comparé à des souris contrôles. Après fécondation, ce différentiel de méthylation n'était plus détectable dans la masse cellulaire interne des blastocystes, suggérant un effacement des DMR acquises en réponse au stress. Néanmoins, après la vague de méthylation *de novo* les chercheurs ont retrouvé pour certaines de ces DMR une différence de méthylation dans la ligne primitive à E7.5 entre les embryons des souris contrôles et des souris soumises à un stress. Ces résultats montrent donc une forme de « mémoire épigénétique », qui permet d'expliquer que certaines DMR soient transmises à la descendance sans résister pour autant à la vague de déméthylation post-fécondation.

Une étude menée par Kremsky et Corces chez la souris propose des mécanismes moléculaires pouvant potentiellement expliquer cette « mémoire » épigénétique. En intégrant des données d'analyse de la méthylation et d'analyse de la chromatine accessible (ATAC-Seq ; « Assay for Transposase-Accessible Chromatin Sequencing »), ils ont montré que le statut de méthylation de 78% de CpG était reprogrammé dans le génome embryonnaire (vague de déméthylation suivie par une méthylation *de novo*). Parmi les CpG conservant un statut de méthylation gamétique chez l'embryon, beaucoup était « occupés » par des facteurs de transcription, qui auraient pu limiter l'accessibilité de ses sites à la machinerie de reprogrammation. Cet article suggère donc que certains facteurs de transcription pourraient constituer un relais afin de transmettre un état de méthylation à travers les générations.

Dans leur ensemble, ces études sont intéressantes et montrent que les régions de méthylome spermatique qui sont transmises à l'embryon pourraient excéder les ~20% de CpG résistants à la déméthylation post-fécondation.

A cela il faut ajouter que les spermatozoïdes apportent d'autres éléments épigénétiques lors de la fécondation : le contenu en sncRNA et les marques post-traductionnelles présentes sur les histones non remplacées par des protamines. Ainsi, caractériser l'épigénome paternel permettrait de s'intéresser aux éléments non génétiques transmis à la descendance.

Certaines marques épigénétiques, pouvant être modulées par des variations environnementales, pourraient ainsi être transmises à la descendance et avoir une influence sur son phénotype. Ces différents aspects s'insèrent directement dans les définitions de la DOHaD (« Developmental Origin of Health and Disease ») et de la POHaD (« Paternal Origin of Health and Disease ») qui sont des recherches s'intéressant à l'impact environnemental maternel ou paternel au cours ou avant le développement de l'embryon sur son phénotype à long terme (Bianco-Miotto *et al.*, 2017; Soubry, 2018). Ces champs de recherche ne sont pas exclusifs à l'épigénétique, mais incluent l'épigénétique comme une des bases moléculaires de ces phénomènes. L'hypothèse de la POHaD postule qu'une modification environnementale peut avoir des conséquences sur la qualité du spermatozoïde,

entrainant un développement embryonnaire non optimal avec un impact à long terme sur le phénotype de la descendance. Il a par exemple été montré que des souris mâles exposées à un régime riche en gras engendraient chez leurs descendants femelles, une altératon de la tolérance au glucose ainsi que des sécrétions insuliniques comparé à des descendants de père ayant eu des régimes normaux. Il a de plus été montré que 642 gènes sont dérégulés dans les cellules béta pancréatiques, montrant donc un effet intergénérationnel du régime (Ng *et al.*, 2010). Néanmoins, ces analyses ne démontrent pas l'origine épigénétique de cette différence. Chez le bovin, une étude s'est concentrée sur la descendance (plus de 6000 filles par taureau) d'une paire de taureaux monozygotiques, donc de même génotype (Liu *et al.*, 2019). Les auteurs ont montré que pour 10 sur 11 caractères de fertilité, les filles du taureau n°1 avaient des performances plus faibles. Les deux taureaux arborant le même génotype, ces différences ne peuvent pas avoir de cause génétique. Les auteurs ont donc analysé le méthylome spermatique de ces taureaux, et ont mis en évidence 528 DMR, qui étaient stables en fonction de l'âge des animaux. Certaines de ces DMR touchent les séquences promotrices de gènes liés à la différenciation sexuelle, au développement gonadique et à des processus reproductifs. Cette étude met donc en évidence une association entre variations du méthylome spermatique et performances de la descendance. Les deux articles cités ci-dessus sont des exemples, ce champ thématique est assez riche et des détails plus complets peuvent se trouver dans ces trois revues (Champroux *et al.*, 2018; Donkin and Barrès, 2018; Illum *et al.*, 2018). Cela suggère donc un lien entre environnement, épigénome spermatique et phénotype de la descendance.

Pour résumer, une partie de l'épigénome spermatique est transmise à l'embryon par la méthylation de l'ADN, les sncRNA, les marques rémanentes d'histones et peut-être d'autres éléments. Ces éléments transmis, bien qu'étant reprogrammés pour certains, sont impliqués dans la physiologie de l'embryon, la trajectoire du développement fœtal et donc conditionne une partie de l'individu en devenir.

Conclusion

L'interrogation posée précédemment sur l'intérêt de l'utilisation de l'épigénétique pour prédire la fertilité peut donc trouver une réponse. En effet, maîtriser l'épigénome paternel, c'est potentiellement maitriser une partie de la variance phénotypique de la descendance, ce qui est donc intéressant dans un contexte agronomique. Aujourd'hui, quelques études menées dans des espèces agronomiques mettent en relation des modifications du méthylome spermatique avec des performances de la progéniture (Braunschweig *et al.*, 2012; Liu *et al.*, 2019; Gross *et al.*, 2020). Néanmoins, le catalogue est loin d'être exhaustif, ce qui n'est pas très étonnant car ces préoccupations sont finalement assez récentes.

Ainsi, si demain dans les systèmes d'élevage un outil d'épigénotypage de routine était disponible, il serait possible de prédire la fertilité mâle mais peut-être également de capter une partie de l'information non génétique transmise à l'embryon, et qui pourrait affecter le devenir à long terme de ce dernier. Cet outil pourrait ouvrir la voie à un grand nombre d'applications, portant sur l'identification de variances phénotypiques imputables à l'épigénome spermatique. Ces informations permettraient peut-être d'identifier des « signatures épigénétiques » favorisant certains caractères d'intérêt comme par exemple la santé des animaux, l'efficience alimentaire, la résistance au stress thermique. S'il existe aujourd'hui un nombre important d'arguments en faveur de transmissions trans/intergénérationnelles d'information épigénétique chez les espèces modèles, les preuves de transmissions épigénétiques impactant quantitativement des caractères agronomiques sont actuellement peu nombreuses. Il est donc difficile de statuer à l'avance si ces recherches aboutiront à des résultats concluants. Cependant, si c'était le cas, les travaux menés pendant ma thèse pourraient s'inscrire dans une perspective de prédiction de la fertilité mâle tout en maximisant la contribution paternelle non génétique à la descendance, favorisant ainsi une sélection épigénomique qui viendrait appuyer la sélection génomique. Tout en constituant une innovation de rupture majeure dans le domaine de l'élevage, une telle démarche pourrait également permettre, au vu de la nombreuse

descendance des taureaux d'IA, de quantifier précisément la magnitude des phénomènes épigénétiques intergénérationnels, et peut-être également d'en explorer les bases scientifiques.

Bibliographie

Abdi, H., Williams, L.J. and Valentin, D. (2013) 'Multiple factor analysis: principal component analysis for multitable and multiblock data sets', *WIREs Computational Statistics*, 5(2), pp. 149–179. Available at: https://doi.org/10.1002/wics.1246.

Agarwal, A. and Allamaneni, S.S.R. (2004) 'The effect of sperm DNA damage on assisted reproduction outcomes. A review', *Minerva Ginecologica*, 56(3), pp. 235–245.

Ajima, R. *et al.* (2008) 'Deficiency of Myo18B in mice results in embryonic lethality with cardiac myofibrillar aberrations', *Genes to Cells*, 13(10), pp. 987–999. Available at: https://doi.org/10.1111/j.1365-2443.2008.01226.x.

Angeloni, A. and Bogdanovic, O. (2021) 'Sequence determinants, function, and evolution of CpG islands', *Biochemical Society Transactions*, 49(3), pp. 1109–1119. Available at: https://doi.org/10.1042/BST20200695.

Aran, D. *et al.* (2011) 'Replication timing-related and gene body-specific methylation of active human genes', *Human Molecular Genetics*, 20(4), pp. 670–680. Available at: https://doi.org/10.1093/hmg/ddq513.

Aravin, A.A. *et al.* (2008) 'A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice', *Molecular Cell*, 31(6), pp. 785–799. Available at: https://doi.org/10.1016/j.molcel.2008.09.003.

Argelaguet, R. *et al.* (2018) 'Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets', *Molecular Systems Biology*, 14(6), p. e8124. Available at: https://doi.org/10.15252/msb.20178124.

Åsenius, F., Danson, A.F. and Marzi, S.J. (2020) 'DNA methylation in human sperm: a systematic review', *Human Reproduction Update*, 26(6), pp. 841–873. Available at: https://doi.org/10.1093/humupd/dmaa025.

Auclair, G. *et al.* (2014) 'Ontogeny of CpG island methylation and specificity of DNMT3 methyltransferases during embryonic development in the mouse', *Genome Biology*, 15(12), p. 545. Available at: https://doi.org/10.1186/s13059-014-0545-5.

Bähler, M. and Rhoads, A. (2002) 'Calmodulin signaling via the IQ motif', *FEBS letters*, 513(1), pp. 107–113. Available at: https://doi.org/10.1016/s0014-5793(01)03239-2.

Banerjee, O., El Ghaoui, L. and d'Aspremont, A. (2008) 'Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data', *Journal of Machine Learning Research*, 9, pp. 485–516.

Bao, J. and Bedford, M.T. (2016) 'Epigenetic regulation of the histone-to-protamine transition during spermiogenesis', *Reproduction (Cambridge, England)*, 151(5), pp. R55-70. Available at: https://doi.org/10.1530/REP-15-0562.

Bao, W., Kojima, K.K. and Kohany, O. (2015) 'Repbase Update, a database of repetitive elements in eukaryotic genomes', *Mobile DNA*, 6, p. 11. Available at: https://doi.org/10.1186/s13100-015-0041-9.

Barbat, A. *et al.* (2010) 'Female Fertility in French Dairy Breeds: Current Situation and Strategies for Improvement', *Journal of Reproduction and Development*, 56, pp. S15–S21. Available at: https://doi.org/10.1262/jrd.1056S15.

Barlow, D.P. *et al.* (1991) 'The mouse insulin-like growth factor type-2 receptor is imprinted and closely linked to the Tme locus', *Nature*, 349(6304), pp. 84–87. Available at: https://doi.org/10.1038/349084a0.

Barzideh, J., Scott, R.J. and Aitken, R.J. (2013) 'Analysis of the global methylation status of human spermatozoa and its association with the tendency of these cells to enter apoptosis', *Andrologia*, 45(6), pp. 424–429. Available at: https://doi.org/10.1111/and.12033.

Bashtrykov, P. *et al.* (2014) 'The UHRF1 protein stimulates the activity and specificity of the maintenance DNA methyltransferase DNMT1 by an allosteric mechanism', *The Journal of Biological Chemistry*, 289(7), pp. 4106–4115. Available at: https://doi.org/10.1074/jbc.M113.528893.

Beaujean, N. *et al.* (2004) 'Effect of Limited DNA Methylation Reprogramming in the Normal Sheep Embryo on Somatic Cell Nuclear Transfer1', *Biology of Reproduction*, 71(1), pp. 185–193. Available at: https://doi.org/10.1095/biolreprod.103.026559.

BEAUJEAN, N. *et al.* (2020) 'L'épigénétique et la construction du phénotype chez le bovin', *INRAE Productions Animales*, 33(2). Available at: https://doi.org/10.20870/productions-animales.2020.33.2.4477.

Bell, J.T. *et al.* (2012) 'Epigenome-Wide Scans Identify Differentially Methylated Regions for Age and Age-Related Phenotypes in a Healthy Ageing Population', *PLoS Genetics*, 8(4), p. e1002629. Available at: https://doi.org/10.1371/journal.pgen.1002629.

Ben Maamar, M. *et al.* (2022) 'Developmental alterations in DNA methylation during gametogenesis from primordial germ cells to sperm', *iScience*, 25(2), p. 103786. Available at: https://doi.org/10.1016/j.isci.2022.103786.

Berry, D.P., Wall, E. and Pryce, J.E. (2014) 'Genetics and genomics of reproductive performance in dairy and beef cattle', *Animal: An International Journal of Animal Bioscience*, 8 Suppl 1, pp. 105–121. Available at: https://doi.org/10.1017/S1751731114000743.

Bianco-Miotto, T. *et al.* (2017) 'Epigenetics and DOHaD: from basics to birth and beyond', *Journal of Developmental Origins of Health and Disease*, 8(5), pp. 513–519. Available at: https://doi.org/10.1017/S2040174417000733.

Blanco, M. and Cocquet, J. (2019) 'Genetic Factors Affecting Sperm Chromatin Structure', *Advances in Experimental Medicine and Biology*, 1166, pp. 1–28. Available at: https://doi.org/10.1007/978-3-030-21664-1_1.

Blythe, M.J. *et al.* (2021) 'LINE-1 transcription in round spermatids is associated with accretion of 5-carboxylcytosine in their open reading frames', *Communications Biology*, 4(1), p. 691. Available at: https://doi.org/10.1038/s42003-021-02217-8.

Boichard, D. *et al.* (2012) 'Genomic selection in French dairy cattle', *Animal Production Science*, 52(3), pp. 115–120. Available at: https://doi.org/10.1071/AN11119.

Borensztein, M. *et al.* (2017) 'Xist-dependent imprinted X inactivation and the early developmental consequences of its failure', *Nature Structural & Molecular Biology*, 24(3), pp. 226–233. Available at: https://doi.org/10.1038/nsmb.3365.

Bourc'his, D. *et al.* (2001) 'Dnmt3L and the establishment of maternal genomic imprints', *Science (New York, N.Y.)*, 294(5551), pp. 2536–2539. Available at: https://doi.org/10.1126/science.1065848.

Bourc'his, D. and Bestor, T.H. (2004) 'Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L', *Nature*, 431(7004), pp. 96–99. Available at: https://doi.org/10.1038/nature02886.

Branton, D. *et al.* (2008) 'The potential and challenges of nanopore sequencing', *Nature Biotechnology*, 26(10), pp. 1146–1153. Available at: https://doi.org/10.1038/nbt.1495.

Braunschweig, M. *et al.* (2012) 'Investigations on transgenerational epigenetic response down the male line in F2 pigs', *PloS One*, 7(2), p. e30583. Available at: https://doi.org/10.1371/journal.pone.0030583.

Breiman, L. (2001) 'Random Forests', *Machine Learning*, 45(1), pp. 5–32. Available at: https://doi.org/10.1023/A:1010933404324.

Breiman, L. *et al.* (2017) *Classification And Regression Trees*. Boca Raton: Routledge. Available at: https://doi.org/10.1201/9781315139470.

Brenet, F. *et al.* (2011) 'DNA methylation of the first exon is tightly linked to transcriptional silencing', *PloS One*, 6(1), p. e14524. Available at: https://doi.org/10.1371/journal.pone.0014524.

Brown, C.J. *et al.* (1991) 'A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome', *Nature*, 349(6304), pp. 38–44. Available at: https://doi.org/10.1038/349038a0.

Brown, C.J. *et al.* (1992) 'The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus', *Cell*, 71(3), pp. 527–542. Available at: https://doi.org/10.1016/0092-8674(92)90520-m.

Camprubí, C. *et al.* (2016) 'Spermatozoa from infertile patients exhibit differences of DNA methylation associated with spermatogenesis-related processes: an array-based analysis', *Reproductive BioMedicine Online*, 33(6), pp. 709–719. Available at: https://doi.org/10.1016/j.rbmo.2016.09.001.

Carmell, M.A. *et al.* (2007) 'MIWI2 Is Essential for Spermatogenesis and Repression of Transposons in the Mouse Male Germline', *Developmental Cell*, 12(4), pp. 503–514. Available at: https://doi.org/10.1016/j.devcel.2007.03.001.

Champroux, A. *et al.* (2018) 'A Decade of Exploring the Mammalian Sperm Epigenome: Paternal Epigenetic and Transgenerational Inheritance', *Frontiers in Cell and Developmental Biology*, 6, p. 50. Available at: https://doi.org/10.3389/fcell.2018.00050.

Chaumeil, J. *et al.* (2006) 'A novel role for Xist RNA in the formation of a repressive nuclear compartment into which genes are recruited when silenced', *Genes & Development*, 20(16), pp. 2223–2237. Available at: https://doi.org/10.1101/gad.380906.

Chen, H. *et al.* (2022) 'Sperm Heterogeneity Accounts for Sperm DNA Methylation Variations Observed in the Caput Epididymis, Independently From DNMT/TET Activities', *Frontiers in Cell and Developmental Biology*, 10, p. 834519. Available at: https://doi.org/10.3389/fcell.2022.834519.

Chen, Q., Yan, W. and Duan, E. (2016) 'Epigenetic inheritance of acquired traits through sperm RNAs and sperm RNA modifications', *Nature Reviews Genetics*, 17(12), pp. 733–743. Available at: https://doi.org/10.1038/nrg.2016.106.

Chiquoine, A.D. (1954) 'The identification, origin, and migration of the primordial germ cells in the mouse embryo', *The Anatomical Record*, 118(2), pp. 135–146. Available at: https://doi.org/10.1002/ar.1091180202.

Chu, C. *et al.* (2017) 'A sequence of 28S rRNA-derived small RNAs is enriched in mature sperm and various somatic tissues and possibly associates with inflammation', *Journal of Molecular Cell Biology*, 9(3), pp. 256–259. Available at: https://doi.org/10.1093/jmcb/mjx016.

Chu, C. *et al.* (2019) 'Epididymal small non-coding RNA studies: progress over the past decade', *Andrology*, 7(5), pp. 681–689. Available at: https://doi.org/10.1111/andr.12639.

Clemson, C.M. *et al.* (1996) 'XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure', *The Journal of Cell Biology*, 132(3), pp. 259–275. Available at: https://doi.org/10.1083/jcb.132.3.259.

Conine, C.C. *et al.* (2018) 'Small RNAs Gained during Epididymal Transit of Sperm Are Essential for Embryonic Development in Mice', *Developmental Cell*, 46(4), pp. 470-480.e3. Available at: https://doi.org/10.1016/j.devcel.2018.06.024.

Costes, V. *et al.* (2022) 'Predicting male fertility from the sperm methylome: application to 120 bulls with hundreds of artificial insemination records', *Clinical Epigenetics*, 14(1), p. 54. Available at: https://doi.org/10.1186/s13148-022-01275-x.

Csankovszki, G., Nagy, A. and Jaenisch, R. (2001) 'Synergism of Xist RNA, DNA methylation, and histone hypoacetylation in maintaining X chromosome inactivation', *The Journal of Cell Biology*, 153(4), pp. 773–784. Available at: https://doi.org/10.1083/jcb.153.4.773.

Czech, B. and Hannon, G.J. (2016) 'One Loop to Rule Them All: The Ping-Pong Cycle and piRNA-Guided Silencing', *Trends in Biochemical Sciences*, 41(4), pp. 324–337. Available at: https://doi.org/10.1016/j.tibs.2015.12.008.

De Jonge, C. (2005) 'Biological basis for human capacitation', *Human Reproduction Update*, 11(3), pp. 205–214. Available at: https://doi.org/10.1093/humupd/dmi010.

Deaton, A.M. and Bird, A. (2011) 'CpG islands and the regulation of transcription', *Genes & Development*, 25(10), pp. 1010–1022. Available at: https://doi.org/10.1101/gad.2037511.

Do, C. *et al.* (2017) 'Genetic-epigenetic interactions in cis: a major focus in the post-GWAS era', *Genome Biology*, 18(1), p. 120. Available at: https://doi.org/10.1186/s13059-017-1250-y.

Dobbs, K.B. *et al.* (2013) 'Dynamics of DNA methylation during early development of the preimplantation bovine embryo', *PloS One*, 8(6), p. e66230. Available at: https://doi.org/10.1371/journal.pone.0066230.

Dong, H. *et al.* (2017) 'Abnormal Methylation of Imprinted Genes and Cigarette Smoking: Assessment of Their Association With the Risk of Male Infertility', *Reproductive Sciences (Thousand Oaks, Calif.)*, 24(1), pp. 114–123. Available at: https://doi.org/10.1177/1933719116650755.

Donkin, I. and Barrès, R. (2018) 'Sperm epigenetics and influence of environmental factors', *Molecular Metabolism*, 14, pp. 1–11. Available at: https://doi.org/10.1016/j.molmet.2018.02.006.

Du, Q. *et al.* (2015) 'Methyl-CpG-binding domain proteins: readers of the epigenome', *Epigenomics*, 7(6), pp. 1051–1073. Available at: https://doi.org/10.2217/epi.15.39.

Eicher, T. *et al.* (2020) 'Metabolomics and Multi-Omics Integration: A Survey of Computational Methods and Resources', *Metabolites*, 10(5). Available at: https://doi.org/10.3390/metabo10050202.

Escofier, B. and Pages, J. (1994) 'Multiple Factor-Analysis (afmult Package)', *Computational Statistics & Data Analysis*, 18(1), pp. 121–140. Available at: https://doi.org/10.1016/0167-9473(94)90135-X.

Evans, J.P. (2002) 'The molecular basis of sperm-oocyte membrane interactions during mammalian fertilization', *Human Reproduction Update*, 8(4), pp. 297–311. Available at: https://doi.org/10.1093/humupd/8.4.297.

Fang, L. *et al.* (2019) 'Integrating Signals from Sperm Methylome Analysis and Genome-Wide Association Study for a Better Understanding of Male Fertility in Cattle', *Epigenomes*, 3(2), p. 10. Available at: https://doi.org/10.3390/epigenomes3020010.

Fellinghauer, B. *et al.* (2013) 'Stable graphical model estimation with Random Forests for discrete, continuous, and mixed variables', *Computational Statistics & Data Analysis*, 64, pp. 132–152. Available at: https://doi.org/10.1016/j.csda.2013.02.022.

Florman, H.M. and Wassarman, P.M. (1985) 'O-linked oligosaccharides of mouse egg ZP3 account for its sperm receptor activity', *Cell*, 41(1), pp. 313–324. Available at: https://doi.org/10.1016/0092-8674(85)90084-4.

Fortes, M.R.S. *et al.* (2013) 'Genomic regions associated with fertility traits in male and female cattle: Advances from microsatellites to high-density chips and beyond', *Animal Reproduction Science*, 141(1), pp. 1–19. Available at: https://doi.org/10.1016/j.anireprosci.2013.07.002.

Foster, J.A. and Gerton, G.L. (2016) 'The Acrosomal Matrix', *Advances in Anatomy, Embryology, and Cell Biology*, 220, pp. 15–33. Available at: https://doi.org/10.1007/978-3-319-30567-7_2.

Friedländer, M.R. *et al.* (2012) 'miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades', *Nucleic Acids Research*, 40(1), pp. 37–52. Available at: https://doi.org/10.1093/nar/gkr688.

Friedman, J., Hastie, T. and Tibshirani, R. (2008) 'Sparse inverse covariance estimation with the graphical lasso', *Biostatistics (Oxford, England)*, 9(3), pp. 432–441. Available at: https://doi.org/10.1093/biostatistics/kxm045.

Friedman, J.H. (2001) 'Greedy Function Approximation: A Gradient Boosting Machine', *The Annals of Statistics*, 29(5), pp. 1189–1232.

Fritz, S. *et al.* (2013) 'Detection of Haplotypes Associated with Prenatal Death in Dairy Cattle and Identification of Deleterious Mutations in GART, SHBG and SLC37A2', *PLoS ONE*, 8(6), p. e65550. Available at: https://doi.org/10.1371/journal.pone.0065550.

García-Ruiz, A. *et al.* (2016) 'Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection', *Proceedings of the National Academy of Sciences of the United States of America*, 113(28), pp. E3995–E4004. Available at: https://doi.org/10.1073/pnas.1519061113.

Gatewood, J.M. *et al.* (1987) 'Sequence-specific packaging of DNA in human sperm chromatin', *Science (New York, N.Y.)*, 236(4804), pp. 962–964. Available at: https://doi.org/10.1126/science.3576213.

Gertz, J. *et al.* (2011) 'Analysis of DNA Methylation in a Three-Generation Family Reveals Widespread Genetic Influence on Epigenetic Regulation', *PLoS Genetics*, 7(8), p. e1002228. Available at: https://doi.org/10.1371/journal.pgen.1002228.

Godmann, M. *et al.* (2007) 'Dynamic regulation of histone H3 methylation at lysine 4 in mammalian spermatogenesis', *Biology of Reproduction*, 77(5), pp. 754–764. Available at: https://doi.org/10.1095/biolreprod.107.062265.

Gross, N. *et al.* (2020) 'The Intergenerational Impacts of Paternal Diet on DNA Methylation and Offspring Phenotypes in Sheep', *Frontiers in Genetics*, 11, p. 597943. Available at: https://doi.org/10.3389/fgene.2020.597943.

Gross, N., Peñagaricano, F. and Khatib, H. (2020) 'Integration of whole-genome DNA methylation data with RNA sequencing data to identify markers for bull fertility', *Animal Genetics*, 51(4), pp. 502–510. Available at: https://doi.org/10.1111/age.12941.

Grundberg, E. *et al.* (2013) 'Global Analysis of DNA Methylation Variation in Adipose Tissue from Twins Reveals Links to Disease-Associated Variants in Distal Regulatory Elements', *American Journal of Human Genetics*, 93(5), pp. 876–890. Available at: https://doi.org/10.1016/j.ajhg.2013.10.004.

Guo, L. *et al.* (2017) 'Sperm-carried RNAs play critical roles in mouse embryonic development', *Oncotarget*, 8(40), pp. 67394–67405. Available at: https://doi.org/10.18632/oncotarget.18672.

Gygi, S.P. *et al.* (1999) 'Correlation between Protein and mRNA Abundance in Yeast', *Molecular and Cellular Biology*, 19(3), pp. 1720–1730.

Ha, M. and Kim, V.N. (2014) 'Regulation of microRNA biogenesis', *Nature Reviews Molecular Cell Biology*, 15(8), pp. 509–524. Available at: https://doi.org/10.1038/nrm3838.

Hackett, J.A. *et al.* (2013) 'Germline DNA Demethylation Dynamics and Imprint Erasure Through 5-Hydroxymethylcytosine', *Science*, 339(6118), pp. 448–452. Available at: https://doi.org/10.1126/science.1229277.

Hammoud, S.S. *et al.* (2009) 'Distinctive chromatin in human sperm packages genes for embryo development', *Nature*, 460(7254), pp. 473–478. Available at: https://doi.org/10.1038/nature08162.

Hammoud, S.S. *et al.* (2011) 'Genome-wide analysis identifies changes in histone retention and epigenetic modifications at developmental and imprinted gene loci in the sperm of infertile men', *Human Reproduction (Oxford, England)*, 26(9), pp. 2558–2569. Available at: https://doi.org/10.1093/humrep/der192.

Hannon, E. *et al.* (2018) 'Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins', *PLoS Genetics*, 14(8), p. e1007544. Available at: https://doi.org/10.1371/journal.pgen.1007544.

Hata, K. *et al.* (2002) 'Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice', *Development (Cambridge, England)*, 129(8), pp. 1983–1993.

Hawk, H.W. (1987) 'Transport and Fate of Spermatozoa After Insemination of Cattle', *Journal of Dairy Science*, 70(7), pp. 1487–1503. Available at: https://doi.org/10.3168/jds.S0022-0302(87)80173-X.

Hazzouri, M. *et al.* (2000) 'Regulated hyperacetylation of core histones during mouse spermatogenesis: involvement of histone deacetylases', *European Journal of Cell Biology*, 79(12), pp. 950–960. Available at: https://doi.org/10.1078/0171-9335-00123.

Hermann, A., Goyal, R. and Jeltsch, A. (2004) 'The Dnmt1 DNA-(cytosine-C5)-methyltransferase methylates DNA processively with high preference for hemimethylated target sites', *The Journal of Biological Chemistry*, 279(46), pp. 48350–48359. Available at: https://doi.org/10.1074/jbc.M403427200.

Holden, S.A. *et al.* (2017) 'Relationship between in vitro sperm functional assessments, seminal plasma composition, and field fertility after AI with either non-sorted or sex-sorted bull semen', *Theriogenology*, 87, pp. 221–228. Available at: https://doi.org/10.1016/j.theriogenology.2016.08.024.

Horsthemke, M. *et al.* (2019) 'A novel isoform of myosin 18A (Myo18Aγ) is an essential sarcomeric protein in mouse heart', *The Journal of Biological Chemistry*, 294(18), pp. 7202–7218. Available at: https://doi.org/10.1074/jbc.RA118.004560.

Horvath, S. (2013) 'DNA methylation age of human tissues and cell types', *Genome Biology*, 14(10), p. 3156. Available at: https://doi.org/10.1186/gb-2013-14-10-r115.

Hotchkiss, R.D. (1948) 'The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography', *The Journal of Biological Chemistry*, 175(1), pp. 315–332.

Hotelling, H. (1992) 'Relations Between Two Sets of Variates', in S. Kotz and N.L. Johnson (eds) *Breakthroughs in Statistics: Methodology and Distribution*. New York, NY: Springer (Springer Series in Statistics), pp. 162–190. Available at: https://doi.org/10.1007/978-1-4612-4380-9_14.

Hua, M. *et al.* (2019) 'Identification of small non-coding RNAs as sperm quality biomarkers for in vitro fertilization', *Cell Discovery*, 5, p. 20. Available at: https://doi.org/10.1038/s41421-019-0087-9.

Huan, T. *et al.* (2019) 'Genome-wide identification of DNA methylation QTLs in whole blood highlights pathways for cardiovascular disease', *Nature Communications*, 10(1), p. 4267. Available at: https://doi.org/10.1038/s41467-019-12228-z.

Huang, D.W., Sherman, B.T. and Lempicki, R.A. (2009) 'Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources', *Nature Protocols*, 4(1), pp. 44–57. Available at: https://doi.org/10.1038/nprot.2008.211.

Hutvágner, G. *et al.* (2001) 'A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA', *Science (New York, N.Y.)*, 293(5531), pp. 834–838. Available at: https://doi.org/10.1126/science.1062961.

Huynh-Thu, V.A. *et al.* (2010) 'Inferring regulatory networks from expression data using tree-based methods', *PloS One*, 5(9), p. e12776. Available at: https://doi.org/10.1371/journal.pone.0012776.

Illum, L.R.H. *et al.* (2018) 'DNA methylation in epigenetic inheritance of metabolic diseases through the male germ line', *Journal of Molecular Endocrinology*, 60(2), pp. R39–R56. Available at: https://doi.org/10.1530/JME-17-0189.

Inoue, A. *et al.* (2011) 'Generation and replication-dependent dilution of 5fC and 5caC during mouse preimplantation development', *Cell Research*, 21(12), pp. 1670–1676. Available at: https://doi.org/10.1038/cr.2011.189.

Inoue, A. and Zhang, Y. (2011) 'Replication-dependent loss of 5-hydroxymethylcytosine in mouse preimplantation embryos', *Science (New York, N.Y.)*, 334(6053), p. 194. Available at: https://doi.org/10.1126/science.1212483.

Inoue, N. *et al.* (2005) 'The immunoglobulin superfamily protein Izumo is required for sperm to fuse with eggs', *Nature*, 434(7030), pp. 234–238. Available at: https://doi.org/10.1038/nature03362.

Iqbal, K. *et al.* (2011) 'Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine', *Proceedings of the National Academy of Sciences of the United States of America*, 108(9), pp. 3642–3647. Available at: https://doi.org/10.1073/pnas.1014033108.

Jachowicz, J.W. *et al.* (2017) 'LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo', *Nature Genetics*, 49(10), pp. 1502–1510. Available at: https://doi.org/10.1038/ng.3945.

Jenkins, T.G. *et al.* (2016) 'Teratozoospermia and asthenozoospermia are associated with specific epigenetic signatures', *Andrology*, 4(5), pp. 843–849. Available at: https://doi.org/10.1111/andr.12231.

Jenkins, T.G., Aston, K.I. and Carrell, D.T. (2018) 'Sperm epigenetics and aging', *Translational Andrology and Urology*, 7(Suppl 3), pp. S328–S335. Available at: https://doi.org/10.21037/tau.2018.06.10.

Jiang, Z. *et al.* (2018) 'DNA methylomes of bovine gametes and in vivo produced preimplantation embryos', *Biology of Reproduction*, 99(5), pp. 949–959. Available at: https://doi.org/10.1093/biolre/ioy138.

Jungwirth, A. *et al.* (2012) 'European Association of Urology Guidelines on Male Infertility: The 2012 Update', *European Urology*, 62(2), pp. 324–332. Available at: https://doi.org/10.1016/j.eururo.2012.04.048.

Kagiwada, S. *et al.* (2013) 'Replication-coupled passive DNA demethylation for the erasure of genome imprints in mice', *The EMBO Journal*, 32(3), pp. 340–353. Available at: https://doi.org/10.1038/emboj.2012.331.

Kaneda, M. *et al.* (2004) 'Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting', *Nature*, 429(6994), pp. 900–903. Available at: https://doi.org/10.1038/nature02633.

Kazazian, H.H. (2004) 'Mobile elements: drivers of genome evolution', *Science (New York, N.Y.)*, 303(5664), pp. 1626–1632. Available at: https://doi.org/10.1126/science.1089670.

Kelsey, G. and Feil, R. (2013) 'New insights into establishment and maintenance of DNA methylation imprints in mammals', *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 368(1609), p. 20110336. Available at: https://doi.org/10.1098/rstb.2011.0336.

Kläver, R. *et al.* (2013) 'DNA methylation in spermatozoa as a prospective marker in andrology', *Andrology*, 1(5), pp. 731–740. Available at: https://doi.org/10.1111/j.2047-2927.2013.00118.x.

Kobayashi, H. *et al.* (2013) 'High-resolution DNA methylome analysis of primordial germ cells identifies gender-specific reprogramming in mice', *Genome Research*, 23(4), pp. 616–627. Available at: https://doi.org/10.1101/gr.148023.112.

Kohli, R.M. and Zhang, Y. (2013) 'TET enzymes, TDG and the dynamics of DNA demethylation', *Nature*, 502(7472), pp. 472–479. Available at: https://doi.org/10.1038/nature12750.

Kohlrausch, F.B. *et al.* (2022) 'Control of LINE-1 Expression Maintains Genome Integrity in Germline and Early Embryo Development', *Reproductive Sciences (Thousand Oaks, Calif.)*, 29(2), pp. 328–340. Available at: https://doi.org/10.1007/s43032-021-00461-1.

Kornberg, R.D. and Lorch, Y. (1999) 'Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome', *Cell*, 98(3), pp. 285–294. Available at: https://doi.org/10.1016/s0092-8674(00)81958-3.

Kremsky, I. and Corces, V.G. (2020) 'Protection from DNA re-methylation by transcription factors in primordial germ cells and pre-implantation embryos can explain trans-generational epigenetic inheritance', *Genome Biology*, 21(1), p. 118. Available at: https://doi.org/10.1186/s13059-020-02036-w.

Kropp, J. *et al.* (2017) 'Male fertility status is associated with DNA methylation signatures in sperm and transcriptomic profiles of bovine preimplantation embryos', *BMC genomics*, 18(1), p. 280. Available at: https://doi.org/10.1186/s12864-017-3673-y.

Krueger, F. and Andrews, S.R. (2011) 'Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications', *Bioinformatics (Oxford, England)*, 27(11), pp. 1571–1572. Available at: https://doi.org/10.1093/bioinformatics/btr167.

Kuramochi-Miyagawa, S. *et al.* (2008) 'DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes', *Genes & Development*, 22(7), pp. 908–917. Available at: https://doi.org/10.1101/gad.1640708.

Kuscu, C. *et al.* (2018) 'tRNA fragments (tRFs) guide Ago to regulate gene expression post-transcriptionally in a Dicer-independent manner', *RNA (New York, N.Y.)*, 24(8), pp. 1093–1105. Available at: https://doi.org/10.1261/rna.066126.118.

Kuster, J.E. *et al.* (1997) 'IAP insertion in the murine LamB3 gene results in junctional epidermolysis bullosa', *Mammalian Genome: Official Journal of the International Mammalian Genome Society*, 8(9), pp. 673–681. Available at: https://doi.org/10.1007/s003359900535.

La Salle, S. *et al.* (2004) 'Windows for sex-specific methylation marked by DNA methyltransferase expression profiles in mouse germ cells', *Developmental Biology*, 268(2), pp. 403–415. Available at: https://doi.org/10.1016/j.ydbio.2003.12.031.

Lambert, M., Benmoussa, A. and Provost, P. (2019) 'Small Non-Coding RNAs Derived from Eukaryotic Ribosomal RNA', *Non-Coding RNA*, 5(1), p. 16. Available at: https://doi.org/10.3390/ncrna5010016.

Lambert, S. *et al.* (2018) 'Spermatozoa DNA methylation patterns differ due to peripubertal age in bulls', *Theriogenology*, 106, pp. 21–29. Available at: https://doi.org/10.1016/j.theriogenology.2017.10.006.

Lane, N. *et al.* (2003) 'Resistance of IAPs to methylation reprogramming may provide a mechanism for epigenetic inheritance in the mouse', *Genesis (New York, N.Y.: 2000)*, 35(2), pp. 88–93. Available at: https://doi.org/10.1002/gene.10168.

Lappalainen, T. and Greally, J.M. (2017) 'Associating cellular epigenetic models with human phenotypes', *Nature Reviews Genetics*, 18(7), pp. 441–451. Available at: https://doi.org/10.1038/nrg.2017.32.

Laqqan, M., Tierling, S., Alkhaled, Y., Porto, C.L., *et al.* (2017) 'Aberrant DNA methylation patterns of human spermatozoa in current smoker males', *Reproductive Toxicology (Elmsford, N.Y.)*, 71, pp. 126–133. Available at: https://doi.org/10.1016/j.reprotox.2017.05.010.

Laqqan, Mohammed *et al.* (2017) 'Alterations in sperm DNA methylation patterns of oligospermic males', *Reproductive Biology*, 17(4), pp. 396–400. Available at: https://doi.org/10.1016/j.repbio.2017.10.007.

Laqqan, M., Tierling, S., Alkhaled, Y., Lo Porto, C., *et al.* (2017) 'Spermatozoa from males with reduced fecundity exhibit differential DNA methylation patterns', *Andrology*, 5(5), pp. 971–978. Available at: https://doi.org/10.1111/andr.12362.

Laqqan, M., Solomayer, E.-F. and Hammadeh, M. (2017) 'Aberrations in sperm DNA methylation patterns are associated with abnormalities in semen parameters of subfertile males', *Reproductive Biology*, 17(3), pp. 246–251. Available at: https://doi.org/10.1016/j.repbio.2017.05.010.

Laqqan, M., Solomayer, E.F. and Hammadeh, M. (2018) 'Association between alterations in DNA methylation level of spermatozoa at CpGs dinucleotide and male subfertility problems', *Andrologia*, 50(1). Available at: https://doi.org/10.1111/and.12832.

Leahy, T. and de Graaf, S. (2012) 'Seminal Plasma and its Effect on Ruminant Spermatozoa During Processing', *Reproduction in Domestic Animals*, 47(s4), pp. 207–213. Available at: https://doi.org/10.1111/j.1439-0531.2012.02077.x.

Lee, J.D. and Hastie, T.J. (2015) 'Learning the Structure of Mixed Graphical Models', *Journal of Computational and Graphical Statistics: A Joint Publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America*, 24(1), pp. 230–253. Available at: https://doi.org/10.1080/10618600.2014.900500.

Lee, S.-M. *et al.* (2017) 'Intragenic CpG islands play important roles in bivalent chromatin assembly of developmental genes', *Proceedings of the National Academy of Sciences of the United States of America*, 114(10), pp. E1885–E1894. Available at: https://doi.org/10.1073/pnas.1613300114.

Lee, Y. *et al.* (2004) 'MicroRNA genes are transcribed by RNA polymerase II', *The EMBO journal*, 23(20), pp. 4051–4060. Available at: https://doi.org/10.1038/sj.emboj.7600385.

Lees-Murdock, D.J., De Felici, M. and Walsh, C.P. (2003) 'Methylation dynamics of repetitive DNA elements in the mouse germ cell lineage', *Genomics*, 82(2), pp. 230–237. Available at: https://doi.org/10.1016/s0888-7543(03)00105-8.

Lehti, M.S. and Sironen, A. (2017) 'Formation and function of sperm tail structures in association with sperm motility defects', *Biology of Reproduction*, 97(4), pp. 522–536. Available at: https://doi.org/10.1093/biolre/iox096.

Leonhardt, H. *et al.* (1992) 'A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei', *Cell*, 71(5), pp. 865–873. Available at: https://doi.org/10.1016/0092-8674(92)90561-p.

Li, E., Bestor, T.H. and Jaenisch, R. (1992) 'Targeted mutation of the DNA methyltransferase gene results in embryonic lethality', *Cell*, 69(6), pp. 915–926. Available at: https://doi.org/10.1016/0092-8674(92)90611-f.

Li, J.-Y. *et al.* (2004) 'Timing of establishment of paternal methylation imprints in the mouse', *Genomics*, 84(6), pp. 952–960. Available at: https://doi.org/10.1016/j.ygeno.2004.08.012.

Li, S., Xu, Z. and Sheng, J. (2018) 'tRNA-Derived Small RNA: A Novel Regulatory Small Non-Coding RNA', *Genes*, 9(5), p. 246. Available at: https://doi.org/10.3390/genes9050246.

Liu, S. *et al.* (2019) 'Divergence Analyses of Sperm DNA Methylomes between Monozygotic Twin AI Bulls', *Epigenomes*, 3(4), p. 21. Available at: https://doi.org/10.3390/epigenomes3040021.

Liu, Y. *et al.* (2012) 'Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data', *Genome Biology*, 13(7), p. R61. Available at: https://doi.org/10.1186/gb-2012-13-7-r61.

Lonergan, P. (2018) 'Review: Historical and futuristic developments in bovine semen technology', *Animal*, 12, pp. s4–s18. Available at: https://doi.org/10.1017/S175173111800071X.

Luger, K. and Richmond, T.J. (1998) 'The histone tails of the nucleosome', *Current Opinion in Genetics & Development*, 8(2), pp. 140–146. Available at: https://doi.org/10.1016/s0959-437x(98)80134-2.

Lund, E. *et al.* (2004) 'Nuclear export of microRNA precursors', *Science (New York, N.Y.)*, 303(5654), pp. 95–98. Available at: https://doi.org/10.1126/science.1090599.

Lyko, F. (2018) 'The DNA methyltransferase family: a versatile toolkit for epigenetic regulation', *Nature Reviews Genetics*, 19(2), pp. 81–92. Available at: https://doi.org/10.1038/nrg.2017.80.

Magnúsdóttir, E. *et al.* (2013) 'A tripartite transcription factor network regulates primordial germ cell specification in mice', *Nature Cell Biology*, 15(8), pp. 905–915. Available at: https://doi.org/10.1038/ncb2798.

Maier, T., Güell, M. and Serrano, L. (2009) 'Correlation of mRNA and protein in complex biological samples', *FEBS Letters*, 583(24), pp. 3966–3973. Available at: https://doi.org/10.1016/j.febslet.2009.10.036.

Marques, C.J. *et al.* (2004) 'Genomic imprinting in disruptive spermatogenesis', *The Lancet*, 363(9422), pp. 1700–1702. Available at: https://doi.org/10.1016/S0140-6736(04)16256-9.

Marques, C.J. *et al.* (2008) 'Abnormal methylation of imprinted genes in human sperm is associated with oligozoospermia', *Molecular Human Reproduction*, 14(2), pp. 67–74. Available at: https://doi.org/10.1093/molehr/gam093.

Mayer, W. *et al.* (2000) 'Demethylation of the zygotic paternal genome', *Nature*, 403(6769), pp. 501–502. Available at: https://doi.org/10.1038/35000656.

McGrath, J. and Solter, D. (1984) 'Completion of mouse embryogenesis requires both the maternal and paternal genomes', *Cell*, 37(1), pp. 179–183. Available at: https://doi.org/10.1016/0092-8674(84)90313-1.

McLachlan, R.I. (2013) 'Approach to the Patient With Oligozoospermia', *The Journal of Clinical Endocrinology & Metabolism*, 98(3), pp. 873–880. Available at: https://doi.org/10.1210/jc.2012-3650.

Meissner, A. *et al.* (2005) 'Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis', *Nucleic Acids Research*, 33(18), pp. 5868–5877. Available at: https://doi.org/10.1093/nar/gki901.

Molaro, A. *et al.* (2011) 'Sperm Methylation Profiles Reveal Features of Epigenetic Inheritance and Evolution in Primates', *Cell*, 146(6), pp. 1029–1041. Available at: https://doi.org/10.1016/j.cell.2011.08.016.

Moore, L.D., Le, T. and Fan, G. (2013) 'DNA Methylation and Its Basic Function', *Neuropsychopharmacology*, 38(1), pp. 23–38. Available at: https://doi.org/10.1038/npp.2012.112.

Mortusewicz, O. *et al.* (2005) 'Recruitment of DNA methyltransferase I to DNA repair sites', *Proceedings of the National Academy of Sciences of the United States of America*, 102(25), pp. 8905–8909. Available at: https://doi.org/10.1073/pnas.0501034102.

Narud, B. *et al.* (2021) 'Sperm chromatin integrity and DNA methylation in Norwegian Red bulls of contrasting fertility', *Molecular Reproduction and Development*, 88(3), pp. 187–200. Available at: https://doi.org/10.1002/mrd.23461.

Neri, F. *et al.* (2017) 'Intragenic DNA methylation prevents spurious transcription initiation', *Nature*, 543(7643), pp. 72–77. Available at: https://doi.org/10.1038/nature21373.

Ng, S.-F. *et al.* (2010) 'Chronic high-fat diet in fathers programs β-cell dysfunction in female rat offspring', *Nature*, 467(7318), pp. 963–966. Available at: https://doi.org/10.1038/nature09491.

Nixon, B. *et al.* (2015) 'The microRNA signature of mouse spermatozoa is substantially modified during epididymal maturation', *Biology of Reproduction*, 93(4), p. 91. Available at: https://doi.org/10.1095/biolreprod.115.132209.

Noblanc, A., Kocer, A. and Drevet, J.R. (2014) 'Recent knowledge concerning mammalian sperm chromatin organization and its potential weaknesses when facing oxidative challenge', *Basic and Clinical Andrology*, 24, p. 6. Available at: https://doi.org/10.1186/2051-4190-24-6.

Norris, D.P., Brockdorff, N. and Rastan, S. (1991) 'Methylation status of CpG-rich islands on active and inactive mouse X chromosomes', *Mammalian Genome: Official Journal of the International Mammalian Genome Society*, 1(2), pp. 78–83. Available at: https://doi.org/10.1007/BF02443782.

Oakes, C.C. *et al.* (2007) 'Developmental acquisition of genome-wide DNA methylation occurs prior to meiosis in male germ cells', *Developmental Biology*, 307(2), pp. 368–379. Available at: https://doi.org/10.1016/j.ydbio.2007.05.002.

Okamura, N. *et al.* (1985) 'Sodium bicarbonate in seminal plasma stimulates the motility of mammalian spermatozoa through direct activation of adenylate cyclase', *The Journal of Biological Chemistry*, 260(17), pp. 9699–9705.

Okano, M. *et al.* (1999) 'DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development', *Cell*, 99(3), pp. 247–257. Available at: https://doi.org/10.1016/s0092-8674(00)81656-6.

Orozco, L.D. *et al.* (2015) 'Epigenome-wide association of liver methylation patterns and complex metabolic traits in mice', *Cell metabolism*, 21(6), pp. 905–917. Available at: https://doi.org/10.1016/j.cmet.2015.04.025.

Ostermeier, G.C. *et al.* (2004) 'Reproductive biology: delivering spermatozoan RNA to the oocyte', *Nature*, 429(6988), p. 154. Available at: https://doi.org/10.1038/429154a.

Ozata, D.M. *et al.* (2019) 'PIWI-interacting RNAs: small RNAs with big functions', *Nature Reviews Genetics*, 20(2), pp. 89–108. Available at: https://doi.org/10.1038/s41576-018-0073-3.

Ozsolak, F. *et al.* (2008) 'Chromatin structure analyses identify miRNA promoters', *Genes & Development*, 22(22), pp. 3172–3183. Available at: https://doi.org/10.1101/gad.1706508.

Pacheco, S.E. *et al.* (2011) 'Integrative DNA methylation and gene expression analyses identify DNA packaging and epigenetic regulatory genes associated with low motility sperm', *PloS One*, 6(6), p. e20280. Available at: https://doi.org/10.1371/journal.pone.0020280.

Peng, H. *et al.* (2012) 'A novel class of tRNA-derived small RNAs extremely enriched in mature mouse sperm', *Cell Research*, 22(11), pp. 1609–1612. Available at: https://doi.org/10.1038/cr.2012.141.

Perrier, J.-P. *et al.* (2018) 'A multi-scale analysis of bull sperm methylome revealed both species peculiarities and conserved tissue-specific features', *BMC Genomics*, 19. Available at: https://doi.org/10.1186/s12864-018-4764-0.

Peters, A.H. *et al.* (2001) 'Loss of the Suv39h histone methyltransferases impairs mammalian heterochromatin and genome stability', *Cell*, 107(3), pp. 323–337. Available at: https://doi.org/10.1016/s0092-8674(01)00542-6.

Phillips, D.M. and Johns, E.W. (1965) 'A FRACTIONATION OF THE HISTONES OF GROUP F2A FROM CALF THYMUS', *The Biochemical Journal*, 94, pp. 127–130. Available at: https://doi.org/10.1042/bj0940127.

Picard, M. *et al.* (2021) 'Integration strategies of multi-omics data for machine learning analysis', *Computational and Structural Biotechnology Journal*, 19, pp. 3735–3746. Available at: https://doi.org/10.1016/j.csbj.2021.06.0302001-0370/.

Pittoggi, C. *et al.* (1999) 'A fraction of mouse sperm chromatin is organized in nucleosomal hypersensitive domains enriched in retroposon DNA', *Journal of Cell Science*, 112 (Pt 20), pp. 3537–3548.

Pizzol, D., Bertoldo, A. and Foresta, C. (2014) 'Male infertility: biomolecular aspects', *Biomolecular Concepts*, 5(6), pp. 449–456. Available at: https://doi.org/10.1515/bmc-2014-0031.

Poplinski, A. *et al.* (2010) 'Idiopathic male infertility is strongly associated with aberrant methylation of MEST and IGF2/H19 ICR1', *International Journal of Andrology*, 33(4), pp. 642–649. Available at: https://doi.org/10.1111/j.1365-2605.2009.01000.x.

Pradhan, S. *et al.* (1999) 'Recombinant human DNA (cytosine-5) methyltransferase. I. Expression, purification, and comparison of de novo and maintenance methylation', *The Journal of Biological Chemistry*, 274(46), pp. 33002–33010. Available at: https://doi.org/10.1074/jbc.274.46.33002.

Rathke, C. *et al.* (2014) 'Chromatin dynamics during spermiogenesis', *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1839(3), pp. 155–168. Available at: https://doi.org/10.1016/j.bbagrm.2013.08.004.

Rezaeian, A., Karimian, M. and Hossienzadeh Colagar, A. (2021) 'Methylation Status of MTHFR Promoter and Oligozoospermia Risk: An Epigenetic Study and in Silico Analysis', *Cell Journal*, 22(4), pp. 482–490. Available at: https://doi.org/10.22074/cellj.2021.6498.

Roosen-Runge, E.C. (1962) 'The process of spermatogenesis in mammals', *Biological Reviews of the Cambridge Philosophical Society*, 37, pp. 343–377. Available at: https://doi.org/10.1111/j.1469-185x.1962.tb01616.x.

Rosenblatt, F. (1958) 'The Perceptron - a Probabilistic Model for Information-Storage and Organization in the Brain', *Psychological Review*, 65(6), pp. 386–408. Available at: https://doi.org/10.1037/h0042519.

Sadakierska-Chudy, A. and Filip, M. (2015) 'A comprehensive view of the epigenetic landscape. Part II: Histone post-translational modification, nucleosome level, and chromatin regulation by ncRNAs', *Neurotoxicity Research*, 27(2), pp. 172–197. Available at: https://doi.org/10.1007/s12640-014-9508-6.

Saito, K. *et al.* (2007) 'Pimet, the Drosophila homolog of HEN1, mediates 2'-O-methylation of Piwi-interacting RNAs at their 3' ends', *Genes & Development*, 21(13), pp. 1603–1608. Available at: https://doi.org/10.1101/gad.1563607.

Santi, D. *et al.* (2017) 'Impairment of sperm DNA methylation in male infertility: a meta-analytic study', *Andrology*, 5(4), pp. 695–703. Available at: https://doi.org/10.1111/andr.12379.

Sasai, N., Nakao, M. and Defossez, P.-A. (2010) 'Sequence-specific recognition of methylated DNA by human zinc-finger proteins', *Nucleic Acids Research*, 38(15), pp. 5015–5022. Available at: https://doi.org/10.1093/nar/gkq280.

Saxonov, S., Berg, P. and Brutlag, D.L. (2006) 'A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters', *Proceedings of the National Academy of Sciences of the United States of America*, 103(5), pp. 1412–1417. Available at: https://doi.org/10.1073/pnas.0510310103.

Scelfo, A. and Fachinetti, D. (2019) 'Keeping the Centromere under Control: A Promising Role for DNA Methylation', *Cells*, 8(8), p. 912. Available at: https://doi.org/10.3390/cells8080912.

Schübeler, D. (2015) 'Function and information content of DNA methylation', *Nature*, 517(7534), pp. 321–326. Available at: https://doi.org/10.1038/nature14192.

Schütte, B. *et al.* (2013) 'Broad DNA methylation changes of spermatogenesis, inflammation and immune response-related genes in a subgroup of sperm samples for assisted reproduction', *Andrology*, 1(6), pp. 822–829. Available at: https://doi.org/10.1111/j.2047-2927.2013.00122.x.

Seiler Vellame, D. *et al.* (2021) 'Characterizing the properties of bisulfite sequencing data: maximizing power and sensitivity to identify between-group differences in DNA methylation', *BMC Genomics*, 22(1), p. 446. Available at: https://doi.org/10.1186/s12864-021-07721-z.

Seisenberger, S. *et al.* (2012) 'The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells', *Molecular Cell*, 48(6), pp. 849–862. Available at: https://doi.org/10.1016/j.molcel.2012.11.001.

Sellem, E. *et al.* (2015) 'Use of combinations of in vitro quality assessments to predict fertility of bovine semen', *Theriogenology*, 84(9), pp. 1447-1454.e5. Available at: https://doi.org/10.1016/j.theriogenology.2015.07.035.

Sellem, E. *et al.* (2020) 'A comprehensive overview of bull sperm-borne small non-coding RNAs and their diversity across breeds', *Epigenetics & Chromatin*, 13(1), p. 19. Available at: https://doi.org/10.1186/s13072-020-00340-0.

Sellem, E. *et al.* (2021) 'Sperm-borne sncRNAs: potential biomarkers for semen fertility?', *Reproduction, Fertility and Development* [Preprint]. Available at: https://doi.org/10.1071/RD21276.

Sharma, U. *et al.* (2016) 'Biogenesis and function of tRNA fragments during sperm maturation and fertilization in mammals', *Science (New York, N.Y.)*, 351(6271), pp. 391–396. Available at: https://doi.org/10.1126/science.aad6780.

Sharma, U. *et al.* (2018) 'Small RNAs are trafficked from the epididymis to developing mammalian sperm', *Developmental cell*, 46(4), pp. 481-494.e6. Available at: https://doi.org/10.1016/j.devcel.2018.06.023.

Shen, R., Olshen, A.B. and Ladanyi, M. (2009) 'Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis', *Bioinformatics*, 25(22), pp. 2906–2912. Available at: https://doi.org/10.1093/bioinformatics/btp543.

Shoji, M. *et al.* (2009) 'The TDRD9-MIWI2 Complex Is Essential for piRNA-Mediated Retrotransposon Silencing in the Mouse Male Germline', *Developmental Cell*, 17(6), pp. 775–787. Available at: https://doi.org/10.1016/j.devcel.2009.10.012.

Soubry, A. (2018) 'POHaD: why we should study future fathers', *Environmental Epigenetics*, 4(2), p. dvy007. Available at: https://doi.org/10.1093/eep/dvy007.

Ss, S. and Aa, P. (2006) 'Sperm transport in the female reproductive tract', *Human reproduction update*, 12(1). Available at: https://doi.org/10.1093/humupd/dmi047.

Staub, C. and Johnson, L. (2018) 'Review: Spermatogenesis in the bull', *Animal: An International Journal of Animal Bioscience*, 12(s1), pp. s27–s35. Available at: https://doi.org/10.1017/S1751731118000435.

Strobl, C. *et al.* (2007) 'Bias in random forest variable importance measures: Illustrations, sources and a solution', *Bmc Bioinformatics*, 8, p. 25. Available at: https://doi.org/10.1186/1471-2105-8-25.

Subramanian, I. *et al.* (2020) 'Multi-omics Data Integration, Interpretation, and Its Application', *Bioinformatics and Biology Insights*, 14, p. 1177932219899051. Available at: https://doi.org/10.1177/1177932219899051.

Sullivan, R. and Mieusset, R. (2016) 'The human epididymis: its function in sperm maturation', *Human Reproduction Update*, 22(5), pp. 574–587. Available at: https://doi.org/10.1093/humupd/dmw015.

Takeda, K. *et al.* (2019) 'Age-related changes in DNA methylation levels at CpG sites in bull spermatozoa and in vitro fertilization-derived blastocyst-stage embryos revealed by combined bisulfite restriction analysis', *The Journal of Reproduction and Development*, 65(4), pp. 305–312. Available at: https://doi.org/10.1262/jrd.2018-146.

Takeda, K. *et al.* (2021) 'Differentially methylated CpG sites related to fertility in Japanese Black bull spermatozoa: epigenetic biomarker candidates to predict sire conception rate', *The Journal of Reproduction and Development* [Preprint]. Available at: https://doi.org/10.1262/jrd.2020-137.

Taudt, A., Colomé-Tatché, M. and Johannes, F. (2016) 'Genetic sources of population epigenomic variation', *Nature Reviews. Genetics*, 17(6), pp. 319–332. Available at: https://doi.org/10.1038/nrg.2016.45.

Taylor, J.F., Schnabel, R.D. and Sutovsky, P. (2018) 'Genomics of Bull Fertility', *Animal : an international journal of animal bioscience*, 12(Suppl 1), pp. s172–s183. Available at: https://doi.org/10.1017/S1751731118000599.

Tian, M. *et al.* (2014) 'Association of DNA methylation and mitochondrial DNA copy number with human semen quality', *Biology of Reproduction*, 91(4), p. 101. Available at: https://doi.org/10.1095/biolreprod.114.122465.

Tini, G. *et al.* (2019) 'Multi-omics integration—a comparison of unsupervised clustering methodologies', *Briefings in Bioinformatics*, 20(4), pp. 1269–1279. Available at: https://doi.org/10.1093/bib/bbx167.

Toubiana, S. and Selig, S. (2020) 'Human subtelomeric DNA methylation: regulation and roles in telomere function', *Current Opinion in Genetics & Development*, 60, pp. 9–16. Available at: https://doi.org/10.1016/j.gde.2020.02.004.

Vagin, V.V. *et al.* (2006) 'A distinct small RNA pathway silences selfish genetic elements in the germline', *Science (New York, N.Y.)*, 313(5785), pp. 320–324. Available at: https://doi.org/10.1126/science.1129333.

Verma, A. *et al.* (2014) 'Genome-wide profiling of sperm DNA methylation in relation to buffalo (Bubalus bubalis) bull fertility', *Theriogenology*, 82(5), pp. 750-759.e1. Available at: https://doi.org/10.1016/j.theriogenology.2014.06.012.

Waddington, C.H. (2012) 'The epigenotype. 1942', *International Journal of Epidemiology*, 41(1), pp. 10–13. Available at: https://doi.org/10.1093/ije/dyr184.

Wang, C. and Lin, H. (2021) 'Roles of piRNAs in transposon and pseudogene regulation of germline mRNAs and lncRNAs', *Genome Biology*, 22, p. 27. Available at: https://doi.org/10.1186/s13059-020-02221-x.

Wang, L. *et al.* (2014) 'Programming and inheritance of parental DNA methylomes in mammals', *Cell*, 157(4), pp. 979–991. Available at: https://doi.org/10.1016/j.cell.2014.04.017.

Wang, Y. *et al.* (2020) 'Both Cauda and Caput Epididymal Sperm Are Capable of Supporting Full-Term Development in FVB and CD-1 Mice', *Developmental Cell*, 55(6), pp. 675–676. Available at: https://doi.org/10.1016/j.devcel.2020.11.022.

Weber, M. *et al.* (2007) 'Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome', *Nature Genetics*, 39(4), pp. 457–466. Available at: https://doi.org/10.1038/ng1990.

Wei, H. *et al.* (2013) 'Profiling and Identification of Small rDNA-Derived RNAs and Their Potential Biological Functions', *PLoS ONE*, 8(2), p. e56842. Available at: https://doi.org/10.1371/journal.pone.0056842.

Werbos, P.J. (1990) 'Backpropagation through time: what it does and how to do it', *Proceedings of the IEEE*, 78(10), pp. 1550–1560. Available at: https://doi.org/10.1109/5.58337.

Wold, S., Sjöström, M. and Eriksson, L. (2001) 'PLS-regression: a basic tool of chemometrics', *Chemometrics and Intelligent Laboratory Systems*, 58(2), pp. 109–130. Available at: https://doi.org/10.1016/S0169-7439(01)00155-1.

Wosnitzer, M.S. and Goldstein, M. (2014) 'Obstructive Azoospermia', *Urologic Clinics of North America*, 41(1), pp. 83–95. Available at: https://doi.org/10.1016/j.ucl.2013.08.013.

Wu, W. *et al.* (2010) 'Idiopathic male infertility is strongly associated with aberrant promoter methylation of methylenetetrahydrofolate reductase (MTHFR)', *PloS One*, 5(11), p. e13884. Available at: https://doi.org/10.1371/journal.pone.0013884.

Wu, X. and Zhang, Y. (2017) 'TET-mediated active DNA demethylation: mechanism, function and beyond', *Nature Reviews. Genetics*, 18(9), pp. 517–534. Available at: https://doi.org/10.1038/nrg.2017.33.

Yamazaki, T. *et al.* (2020) 'Editing DNA Methylation in Mammalian Embryos', *International Journal of Molecular Sciences*, 21(2), p. E637. Available at: https://doi.org/10.3390/ijms21020637.

Yoder, J.A., Walsh, C.P. and Bestor, T.H. (1997) 'Cytosine methylation and the ecology of intragenomic parasites', *Trends in genetics: TIG*, 13(8), pp. 335–340. Available at: https://doi.org/10.1016/s0168-9525(97)01181-5.

Zealy, R.W. *et al.* (2017) 'microRNA-binding proteins: specificity and function', *Wiley interdisciplinary reviews. RNA*, 8(5). Available at: https://doi.org/10.1002/wrna.1414.

Zheng, X. *et al.* (2021) 'Sperm epigenetic alterations contribute to inter- and transgenerational effects of paternal exposure to long-term psychological stress via evading offspring embryonic reprogramming', *Cell Discovery*, 7, p. 101. Available at: https://doi.org/10.1038/s41421-021-00343-5.

Annexes

Annexe 1 : Taureaux et échantillons de la cohorte longitudinale « Age »

Nom_taureau	Date naissance	Age collecte	Nb_ejaculat	Dose_par_ejaculat
GALOP	23/03/2011	13m	2	4;4
GALOP	23/03/2011	15m	3	3;3;3
GALOP	23/03/2011	17m	2	4;5
GALOP	23/03/2011	19m	1	9
GECKO P RF	04/03/2011	13m	1	8
GECKO P RF	04/03/2011	15m	2	2;2
GECKO P RF	04/03/2011	17m	2	5;4
GECKO P RF	04/03/2011	19m	1	9
GLADIATOR	30/04/2011	13m	3	3;3;3
GLADIATOR	30/04/2011	15m	2	3;3
GLADIATOR	30/04/2011	17m	3	3;3;3
GLADIATOR	30/04/2011	19m	3	3;3;3
GLOUTON	12/05/2011	13m	2	4;4
GLOUTON	12/05/2011	15m	3	3;3;3
GLOUTON	12/05/2011	17m	3	1;5;3
GLOUTON	12/05/2011	19m	2	4;5
GOODTYPE	17/08/2011	13m	1	8
GOODTYPE	17/08/2011	15m	4	2;3;2;2
GOODTYPE	17/08/2011	17m	4	2;3;2;2
GOODTYPE	17/08/2011	19m	3	3;3;3
GOVOU	22/08/2011	13m	3	3;3;3
GOVOU	22/08/2011	15m	2	6;3
GOVOU	22/08/2011	17m	3	3;3;3
GOVOU	22/08/2011	19m	2	4;5
HOLLISTER	03/04/2012	13m	3	3;3;3
HOLLISTER	03/04/2012	15m	2	6;3
HOLLISTER	03/04/2012	17m	3	3;3;3
HOLLISTER	03/04/2012	19m	2	4;5
JUCKY	15/11/2014	13m	1	4
JUCKY	15/11/2014	15m	1	8
JUCKY	15/11/2014	17m	1	9
JUCKY	15/11/2014	19m	1	9

Annexe 2 Taureaux et échantillons de la cohorte « déviation »

Taureaux	Date naissance	Première éjac	Dernière éjac	Groupe fertilité	Statut	Nombre éjaculat	Paillette par éjaculat	Ferti pondéré
GALOP	23/03/2011	20/08/2012	30/08/2012	Fertile	Deviant	2;	5;4	406.8
GALOP	23/03/2011	09/08/2012	27/08/2012	Fertile	Cohérent	2;	4;5	462
GAYOLI	02/05/2011	09/10/2012	26/10/2012	Fertile	Cohérent	2;	3;4	451.4
GAYOLI	02/05/2011	16/10/2012	19/10/2012	Fertile	Déviant	2;	5;4	358.4
GELIZAT	06/12/2011	09/07/2013	09/07/2013	Fertile	Déviant	1;	9	413.9
GELIZAT	06/12/2011	12/07/2013	19/07/2013	Fertile	Cohérent	2;	4;4	446.1
GEMO RF	09/05/2011	20/09/2012	20/09/2012	Subfertile	Cohérent	1;	9	521.8
GEMO RF	09/05/2011	24/09/2012	24/09/2012	Subfertile	Déviant	1;	8	609.6
GIAGI	04/09/2011	28/01/2013	31/01/2013	Fertile	Cohérent	2;	4;4	518.8
GIAGI	04/09/2011	21/01/2013	21/01/2013	Fertile	Déviant	1;	7	419.2
GIMEL	24/03/2011	03/09/2012	03/09/2012	Fertile	Déviant	1;	9	397.1
GIMEL	24/03/2011	06/09/2012	10/09/2012	Fertile	Cohérent	2;	4;5	436.8
GLADIATOR	30/04/2011	13/09/2012	24/09/2012	Fertile	Cohérent	3;	3;3;3	504.7
GLADIATOR	30/04/2011	20/09/2012	20/09/2012	Fertile	Déviant	1;	9	385.6
GLOUTON	12/05/2011	06/11/2012	09/11/2012	Subfertile	Déviant	2;	4;5	551.4
GLOUTON	12/05/2011	02/11/2012	16/11/2012	Subfertile	Cohérent	3;	3;3;3	425.4
GOLESNIL	27/05/2011	19/11/2012	26/11/2012	Fertile	Déviant	2;	5;4	437.6
GOLESNIL	27/05/2011	22/11/2012	29/11/2012	Fertile	Cohérent	2;	4;5	534.7
GOODTYPE	17/08/2011	27/02/2013	27/02/2013	Fertile	Déviant	1;	8;	382.4
GOODTYPE	17/08/2011	15/03/2013	29/03/2013	Fertile	Cohérent	2;	5;4	479.3
GOVOU	22/08/2011	10/01/2013	24/01/2013	Fertile	Déviant	2;	4;5	323.6
GOVOU	22/08/2011	07/01/2013	14/01/2013	Fertile	Cohérent	2;	4;4	417.2
GRANIT	19/02/2011	24/09/2012	24/09/2012	Fertile	Déviant	1;	8	395.4
GRANIT	19/02/2011	20/09/2012	01/10/2012	Fertile	Cohérent	2;	4;5	566.2
GRAPON P	31/07/2011	14/12/2012	14/12/2012	Fertile	Déviant	1;	9	349.3
GRAPON P	31/07/2011	21/12/2012	26/12/2012	Fertile	Cohérent	3;	3;3;3	490.7
GRINJTON	03/03/2011	24/07/2012	24/07/2012	Subfertile	Déviant	1;	8	536
GRINJTON	03/03/2011	03/08/2012	10/08/2012	Subfertile	Cohérent	2;	5;4	455.6
GUESS	29/06/2011	31/12/2012	11/01/2013	Subfertile	Cohérent	2;	5;4	394.3

GUESS	29/06/2011	02/01/2013	02/01/2013	Subfertile	Déviant	1;	8	479.3
HDMI	27/04/2012	10/09/2013	01/10/2013	Subfertile	Cohérent	2;	1;7	NA
HDMI	27/04/2012	27/09/2013	27/09/2013	Subfertile	Déviant	1;	8	402.6
HIBISCUS	16/01/2012	01/07/2013	01/07/2013	Fertile	Cohérent	1;	8	474.7
HIBISCUS	16/01/2012	22/07/2013	22/07/2013	Fertile	Déviant	1;	9	301.4
HOLLISTER	03/04/2012	13/09/2013	13/09/2013	Subfertile	Déviant	1;	8	532.1
HOLLISTER	03/04/2012	03/09/2013	10/09/2013	Subfertile	Cohérent	3;	1;3;4	408.4
JENIX	19/04/2014	21/09/2015	21/09/2015	Fertile	Déviant	1;	9	380.7
JENIX	19/04/2014	07/09/2015	07/09/2015	Fertile	Cohérent	1;	8	460.9
JOCKER	09/01/2014	23/06/2015	23/06/2015	Fertile	Déviant	1;	9	314.8
JOCKER	09/01/2014	26/05/2015	26/05/2015	Fertile	Cohérent	NA	NA	422.3

Annexe 3 : Figures supplémentaire et table supplémentaire de l'article 1

Sample ID	Fertillity	Semen collection center	Cohort	Bull's birth	Nb. of ejaculates per sample	Semen collection date (earliest ejaculate in sample)	Semen collection date (latest ejaculate in sample)	Nb. of straws per ejaculate	Sample preparation batch	Library preparation batch	Corrected NRR 56 (%)	Corrected SCR (%)	Nb. of AIs for corrected NRR 56 evaluation	Nb. of AIs for corrected SCR evaluation
1_1	Subfertile	Center 1	Main	5-Mar-13	3	25-Aug-14	4-Sep-14	3;3;3	1	1	0.4600633	3.9144488	490	845
1_10	Fertile	Center 1	Main	28-Sep-13	4	17-Mar-15	28-Apr-15	2;3;2;2	1	1	4.0153319	-0.0599618	660	1187
1_11	Fertile	Center 1	Main	15-Sep-12	3	18-Feb-14	11-Apr-14	3;3;3	1	2	3.8310095	4.0729941	736	1167
1_12	Fertile	Center 1	Main	5-Sep-14	2	2-Feb-16	9-Feb-16	5;4	1	1	3.6574629	3.9237761	327	586
1_2	Fertile	Center 1	Main	13-Sep-13	3	17-Feb-15	3-Mar-15	3;3;3	1	1	6.181468	4.7744555	396	716
1_3	Subfertile	Center 1	Main	22-Oct-11	4	25-Mar-13	8-Apr-13	2;2;2;3	1	1	-0.1804805	2.5866731	442	707
1_4	Fertile	Center 1	Main	14-Jan-14	3	30-Jun-15	4-Aug-15	3;3;3	1	1	4.8306795	3.5567341	661	1077
1_5	Fertile	Center 1	Main	23-Mar-11	3	25-Sep-12	11-Oct-12	3;3;3	1	1	3.427486	2.3195227	583	969
1_6	Subfertile	Center 1	Main	16-Sep-13	3	10-Mar-15	24-Mar-15	3;3;3	1	1	0.2395069	0.100693	439	817
1_7	Fertile	Center 1	Main	22-Jun-13	5	4-Dec-14	20-Jan-15	2;2;2;2;2	1	1	3.1035796	3.5407255	297	537
1_8	Fertile	Center 1	Main	24-Dec-13	4	30-Jun-15	21-Jul-15	2;2;3;2	1	1	2.9644213	3.9997588	584	917
1_9	Fertile	Center 1	Main	15-Apr-12	3	7-Oct-13	17-Oct-13	3;3;3	1	1	3.9146885	3.8400145	385	622
11_10	Subfertile	Center 2	Main	NA	5	NA	NA	2;2;2;2;2	7	3	0.3112622	0.7701077	246	367
11_11	Subfertile	Center 2	Main	NA	4	NA	NA	2;2;2;3	7	3	-0.4039487	0.9723555	113	220
11_12	Fertile	Center 2	Main	NA	4	NA	NA	2;2;2;3	7	3	4.1732793	7.8871261	197	274
11_13	Fertile	Center 2	Main	NA	5	NA	NA	2;2;2;2;2	7	4	4.5280958	3.9800615	390	632
11_14	Fertile	Center 2	Main	NA	5	NA	NA	2;2;2;2;2	7	5	4.2427517	1.8255566	203	329
11_3	Fertile	Center 2	Main	NA	5	NA	NA	2;2;2;2;2	7	3	5.8050963	3.4046006	275	420
11_4	Fertile	Center 2	Main	NA	5	NA	NA	2;2;2;2;2	7	3	3.9954768	5.8801801	220	328
11_5	Fertile	Center 2	Main	NA	5	NA	NA	2;2;2;2;2	7	3	4.8201802	5.6367058	445	651
11_6	Subfertile	Center 2	Main	NA	5	NA	NA	2;2;2;2;2	7	3	-0.1502004	-1.4458358	301	427

11_7	Fertile	Center 2	Main	3-Sep-11	5	NA	NA	2;2;2;2;2	7	3	2.8955442	2.1563071	215	332
2_1	Fertile	Center 1	Main	16-Sep-14	4	16-Feb-16	22-Mar-16	3;2;2;2	2	6	5.3571503	6.935251	485	809
2_10	Subfertile	Center 1	Main	31-Jan-14	2	30-Jun-15	7-Jul-15	4;5	2	6	-1.1638841	-0.926793	419	757
2_11	Subfertile	Center 1	Main	4-Oct-11	4	18-Mar-13	4-Apr-13	2;2;2;2	2	6	-0.2277923	0.6744516	392	657
2_12	Fertile	Center 1	Main	21-Aug-14	4	16-Feb-16	10-Mar-16	2;3;2;2	2	6	5.2007926	4.9154027	576	1007
2_13	Subfertile	Center 1	Main	20-Jun-14	2	8-Dec-15	19-Jan-16	4;5	2	7	-1.3768726	-1.029771	280	545
2_14	Fertile	Center 1	Main	23-Dec-12	3	23-May-14	3-Jun-14	3;3;3	2	7	3.5487942	4.4799653	413	745
2_15	Fertile	Center 1	Main	25-Aug-11	4	7-Mar-13	21-Mar-13	2;2;3;2	2	7	4.2153489	3.4027881	477	783
2_16	Fertile	Center 1	Main	11-Jul-14	4	5-Jan-16	9-Feb-16	2;3;2;2	2	7	3.8174658	3.6556995	478	840
2_17	Fertile	Center 1	Main	4-Jan-11	4	4-Jun-12	25-Jun-12	3;2;2;2	2	7	4.8666797	4.1782966	2709	4401
2_18	Subfertile	Center 1	Main	27-Jul-12	3	30-Dec-13	3-Feb-14	3;3;3	2	7	0.3289761	0.3266043	125	196
2_19	Fertile	Center 1	Main	26-Nov-13	3	12-May-15	2-Jun-15	3;3;3	2	7	5.2557289	4.2066853	792	1298
2_2	Fertile	Center 1	Main	6-Jun-12	2	18-Nov-13	26-Nov-13	5;4	2	6	7.2494533	4.2705775	655	1292
2_20	Subfertile	Center 1	Main	19-Aug-11	5	21-Feb-13	7-Mar-13	2;2;2;2;2	2	7	-0.6538384	-2.1760461	568	1069
2_21	Subfertile	Center 1	Main	1-Jan-13	3	3-Jun-14	17-Jul-14	3;3;3	2	7	-1.0793466	-1.3414277	192	338
2_22	Fertile	Center 1	Main	Year 2012	4	23-Jan-14	3-Feb-14	2;2;2;3	2	7	3.2531326	1.994432	424	809
2_23	Subfertile	Center 1	Main	20-Aug-11	4	28-Jan-13	18-Mar-13	2;2;2;3	2	7	-1.9316502	-1.1762452	297	516
2_24	Fertile	Center 1	Main	9-Sep-14	4	9-Feb-16	29-Mar-16	2;2;2;3	2	2	4.4337319	4.0724358	527	957
2_3	Fertile	Center 1	Main	22-Aug-12	3	3-Feb-14	14-Feb-14	3;3;3	2	6	4.5050587	3.8296109	355	620
2_4	Subfertile	Center 1	Main	18-Oct-12	4	11-Apr-14	29-Apr-14	2;2;3;2	2	6	-0.133438	3.284048	556	963
2_5	Subfertile	Center 1	Main	27-Mar-14	3	6-Oct-15	27-Oct-15	3;3;3	2	6	-1.1677874	-1.8360664	422	810
2_6	Fertile	Center 1	Main	23-Sep-14	2	23-Feb-16	8-Mar-16	4;5	2	6	3.2563328	2.4524331	602	1095
2_7	Subfertile	Center 1	Main	8-Jul-14	3	15-Dec-15	2-Feb-16	3;3;3	2	6	-0.6777061	-0.8608187	337	570
2_8	Subfertile	Center 1	Main	30-Oct-13	4	12-May-15	2-Jun-15	2;3;2;2	2	6	0.2057988	2.3634118	559	919
2_9	Fertile	Center 1	Main	22-May-13	3	27-Oct-14	1-Dec-14	3;3;3	2	6	4.4703534	4.9289536	380	708
3_1	Subfertile	Center 2	Main	NA	5	NA	NA	2;2;2;2;2	3	8	-2.3791435	1.9973178	234	370
3_10	Subfertile	Center 1	Main	12-Oct-12	5	8-Apr-14	24-Apr-14	2;2;2;2;2	3	8	0.5038552	-0.3975196	392	679
3_11	Fertile	Center 1	Main	13-Feb-13	3	19-Aug-14	25-Aug-14	3;3;3	3	8	2.8666898	3.7918781	1863	2864
3_12	Subfertile	Center 1	Main	16-Jul-13	3	15-Dec-14	5-Jan-15	3;3;3	3	8	-0.9095316	-1.4054665	791	1418
3_13	Fertile	Center 1	Main	6-Mar-13	3	2-Sep-14	8-Sep-14	3;3;3	3	9	3.4925679	3.7892552	755	1306

3_14	Subfertile	Center 1	Main	18-Sep-13	4	9-Mar-15	30-Mar-15	2;2;2;3	3	9	-0.1803258	0.4190544	572	1078
3_15	Subfertile	Center 1	Main	13-Sep-14	4	1-Mar-16	5-Apr-16	2;2;3;2	3	9	0.4422121	0.3177764	458	868
3_16	Fertile	Center 1	Main	21-Sep-12	3	18-Mar-14	24-Mar-14	3;3;3	3	9	5.6426633	4.5093829	826	1314
3_17	Fertile	Center 1	Main	6-May-12	5	14-Oct-13	30-Oct-13	2;2;2;2;2	3	9	3.4791391	1.4742869	683	1317
3_18	Subfertile	Center 1	Main	10-Aug-12	3	17-Feb-14	6-Mar-14	3;3;3	3	9	-0.7717208	-2.4233806	735	1282
3_19	Fertile	Center 1	Main	1-Sep-12	4	18-Feb-14	11-Mar-14	2;2;2;3	3	9	4.8824005	2.1024077	597	1031
3_2	Fertile	Center 2	Main	NA	5	NA	NA	2;2;2;2;2	3	8	3.0560628	2.2317876	532	831
3_20	Subfertile	Center 1	Main	10-Feb-13	4	21-Jul-14	4-Aug-14	3;2;2;2	3	9	0.7923267	-0.0684944	509	943
3_21	Subfertile	Center 1	Main	8-Mar-12	4	17-Sep-13	3-Oct-13	3;2;2;2	3	9	-1.0321443	0.9334126	415	705
3_22	Fertile	Center 1	Main	26-Aug-12	3	23-Jan-14	30-Jan-14	3;3;3	3	9	4.0946225	2.4252706	504	925
3_23	Subfertile	Center 1	Main	5-Feb-14	3	6-Jul-15	16-Jul-15	3;3;3	3	9	-0.3561891	-0.4033767	451	739
3_24	Fertile	Center 1	Main	2-Apr-14	4	8-Sep-15	17-Sep-15	2;2;3;2	3	9	5.7614911	4.6013836	645	1082
3_3	Subfertile	Center 2	Main	NA	5	NA	NA	2;2;2;2;2	3	8	-3.13448121	-0.09837011	183	286
3_4	Subfertile	Center 2	Main	16-Sep-11	5	NA	NA	2;2;2;2;2	3	8	0.013371	1.3369414	198	306
3_5	Fertile	Center 1	Main	9-Dec-12	3	25-Jun-14	9-Jul-14	3;3;3	3	8	3.3820528	6.2193696	912	1516
3_6	Fertile	Center 1	Main	8-Nov-12	4	13-May-14	2-Jun-14	2;3;2;2	3	8	3.0494051	2.1555864	1098	1774
3_7	Subfertile	Center 1	Main	23-Dec-11	3	27-May-13	13-Jun-13	3;3;3	3	8	-0.7762574	1.6326062	342	592
3_8	Fertile	Center 1	Main	4-Feb-14	3	9-Jul-15	20-Jul-15	3;3;3	3	8	3.8331718	4.233723	390	661
3_9	Fertile	Center 1	Main	18-Sep-12	3	1-Apr-14	9-Apr-14	3;3;3	3	8	4.7808915	3.8326111	814	1337
5_20	Subfertile	Center 2	Main	NA	3	NA	NA	3;3;3	4	10	-4.086557	-7.9178829	440	734
5_21	Fertile	Center 2	Main	9-Aug-11	5	NA	NA	2;2;1;2;2	4	10	4.6475368	3.6533773	876	1304
5_22	Subfertile	Center 2	Main	NA	5	NA	NA	2;2;2;2;2	4	10	-1.5620337	-0.8110784	700	1111
5_23	Fertile	Center 2	Main	NA	5	NA	NA	2;2;2;2;2	4	10	5.3172282	2.202831	711	1018
5_24	Fertile	Center 2	Main	NA	5	NA	NA	2;2;2;2;2	4	10	5.0459107	4.1124419	397	581
6_1	Fertile	Center 2	Main	25-Aug-11	5	NA	NA	2;2;2;2;2	5	11	3.0391922	4.0603409	521	797
6_10	Subfertile	Center 2	Main	NA	5	NA	NA	2;2;2;2;2	5	11	-0.0145814	1.1673761	527	791
6_11	Fertile	Center 2	Main	NA	5	NA	NA	2;2;2;2;2	5	11	2.8746136	0.8147103	284	435
6_12	Subfertile	Center 2	Main	NA	5	NA	NA	2;2;2;2;2	5	11	0.3183183	0.6784786	312	528
6_13	Subfertile	Center 2	Main	NA	5	NA	NA	2;2;2;2;2	5	5	-0.0183553	0.0610768	576	948
6_14	Subfertile	Center 2	Main	NA	5	NA	NA	2;2;2;2;2	5	5	0.4413036	3.0209727	224	377

6_15	Fertile	Center 2	Main	18-Oct-11	5	NA	NA	2;2;2;2;2	5	5	4.1563817	1.9340247	209	299
6_16	Subfertile	Center 2	Main	NA	3	NA	NA	3;3;3	5	5	-0.6516063	1.0737945	224	350
6_17	Subfertile	Center 2	Main	NA	4	NA	NA	3;2;2;2	5	5	-0.0350394	-0.1384725	286	454
6_18	Fertile	Center 2	Main	NA	5	NA	NA	2;2;2;2;2	5	5	5.6520484	2.1176092	341	565
6_2	Fertile	Center 2	Main	NA	3	NA	NA	3;3;3	5	11	4.8692403	6.2274258	514	799
6_20	Subfertile	Center 2	Main	NA	5	NA	NA	2;2;2;2;2	5	5	0.5677751	3.084497	576	886
6_21	Fertile	Center 2	Main	NA	5	NA	NA	2;2;2;2;2	5	5	3.0018095	1.5103413	257	388
6_22	Subfertile	Center 2	Main	NA	5	NA	NA	2;2;2;2;2	5	5	-1.2408184	4.3600725	286	459
6_23	Fertile	Center 2	Main	NA	5	NA	NA	2;2;2;2;2	5	5	2.9971303	2.7917583	172	281
6_24	Fertile	Center 2	Main	28-Sep-11	5	NA	NA	2;2;2;2;2	5	5	3.5168226	6.1489093	183	278
6_3	Subfertile	Center 2	Main	NA	5	NA	NA	2;2;2;2;2	5	11	-0.0145856	1.159743	332	524
6_4	Subfertile	Center 2	Main	23-Dec-11	5	NA	NA	2;2;2;2;2	5	11	0.0343244	0.7913551	427	761
6_5	Fertile	Center 2	Main	NA	5	NA	NA	2;2;2;2;2	5	11	3.5939122	1.7630338	326	487
6_6	Fertile	Center 2	Main	NA	5	NA	NA	2;2;2;2;2	5	11	2.7748776	4.8049995	393	572
6_7	Subfertile	Center 2	Main	NA	5	NA	NA	2;2;2;2;2	5	11	-1.9017912	-1.9363667	196	299
6_8	Fertile	Center 2	Main	NA	5	NA	NA	2;2;2;2;2	5	11	3.8538926	2.2474919	222	315
6_9	Subfertile	Center 2	Main	NA	5	NA	NA	2;2;2;2;2	5	11	0.7556521	1.078474	143	232
7_15	Fertile	Center 2	Main	13-Jan-11	5	NA	NA	2;2;2;2;2	6	12	3.3776181	4.2390802	796	1175
7_20	Fertile	Center 2	Main	1-Nov-11	5	NA	NA	2;2;2;2;2	6	12	5.7043449	4.8471158	639	1080
sqm_26	Fertile		Independent	22-Jul-12	1	5-Nov-13	5-Nov-13	1	8	13	2.0457451	2.5215442	520	921
sqm_30	Fertile		Independent	17-Dec-10	1	16-Jul-12	16-Jul-12	1	8	13	2.8336185	3.0406457	2739	4368
sqm_32	Fertile		Independent	5-Jul-12	1	18-Dec-13	18-Dec-13	1	8	14	2.8234845	2.0730345	492	939
sqm_33	Fertile		Independent	8-Dec-11	1	14-May-13	14-May-13	1	8	15	1.366466	2.3939359	400	731
sqm_35	Fertile		Independent	16-Jul-10	1	25-Apr-13	25-Apr-13	1	8	13	3.8404689	5.0669376	4162	6634
sqm_38	Fertile		Independent	16-Jul-12	1	4-Dec-13	4-Dec-13	1	8	13	3.7340683	4.6254743	389	683
sqm_44	Fertile		Independent	30-Jan-12	1	7-May-13	7-May-13	1	9	14	1.2873339	-0.2289464	2235	3552
sqm_49	Fertile		Independent	6-Jan-12	1	23-Jul-13	23-Jul-13	1	9	13	2.4337992	3.1322356	649	1085
sqm_50	Fertile		Independent	6-Mar-10	1	17-Nov-11	17-Nov-11	1	9	13	3.1772759	4.6462072	5664	8939
sqm_55	Fertile		Independent	14-Jun-12	1	23-Dec-13	23-Dec-13	1	10	15	3.5595777	1.7493968	468	815
sqm_71	Subfertile		Independent	17-Dec-10	1	14-Oct-13	14-Oct-13	1	11	13	-2.5942563	-3.8779058	4859	7925

sqm_78	Fertile		Independent	2-Oct-10	1	16-Sep-13	16-Sep-13	1	11	14	3.9168304	3.5655842	3874	6227
sqm_79	Fertile		Independent	26-May-10	1	10-Apr-13	10-Apr-13	1	11	14	2.1729775	1.4149763	7151	11459
sqm_80	Fertile		Independent	15-Sep-09	1	28-Jun-12	28-Jun-12	1	12	16	3.0009798	3.2753442	1789	3003
sqm_81	Fertile		Independent	9-Aug-11	1	26-Jun-13	26-Jun-13	1	12	13	2.9381626	3.4580657	5937	8913
sqm_82	Fertile		Independent	19-Sep-09	1	20-Dec-12	20-Dec-12	1	12	13	2.2714799	1.9538168	10558	16737
sqm_85	Fertile		Independent	6-Jul-12	1	19-Nov-13	19-Nov-13	1	12	15	3.5368386	2.2091791	13525	19610
sqm_9	Subfertile		Independent	20-Mar-11	1	30-Jul-12	30-Jul-12	1	13	13	-1.2336685	0.7476829	740	1195
sqm_90	Subfertile		Independent	12-Jan-12	1	20-Jun-13	20-Jun-13	1	12	15	-0.3600747	0.4149555	969	1617
sqm_92	Subfertile		Independent	7-Mar-12	1	6-Aug-13	6-Aug-13	1	14	14	-0.2934691	3.9229885	446	783

ADDITIONAL DATA FILE

Predicting male fertility from the sperm methylome: application to 120 bulls with hundreds of artificial insemination records

Valentin Costes[1,2,3,4], Aurélie Chaulot-Talmon[1,2], Eli Sellem[1,2,3], Jean-Philippe Perrier[1,2], Anne Aubert-Frambourg[1,2], Luc Jouneau[1,2], Charline Pontlevoy[1,2], Chris Hozé[3,4], Sébastien Fritz[3,4], Mekki Boussaha[4], Chrystelle Le Danvic[3], Marie-Pierre Sanchez[4], Didier Boichard[4], Laurent Schibler[3], Hélène Jammes[1,2], Florence Jaffrézic[4], Hélène Kiefer[1,2]*

[1]Université Paris-Saclay, UVSQ, INRAE, BREED, 78350 Jouy-en-Josas, France.

[2]Ecole Nationale Vétérinaire d'Alfort, BREED, 94700, Maisons-Alfort, France

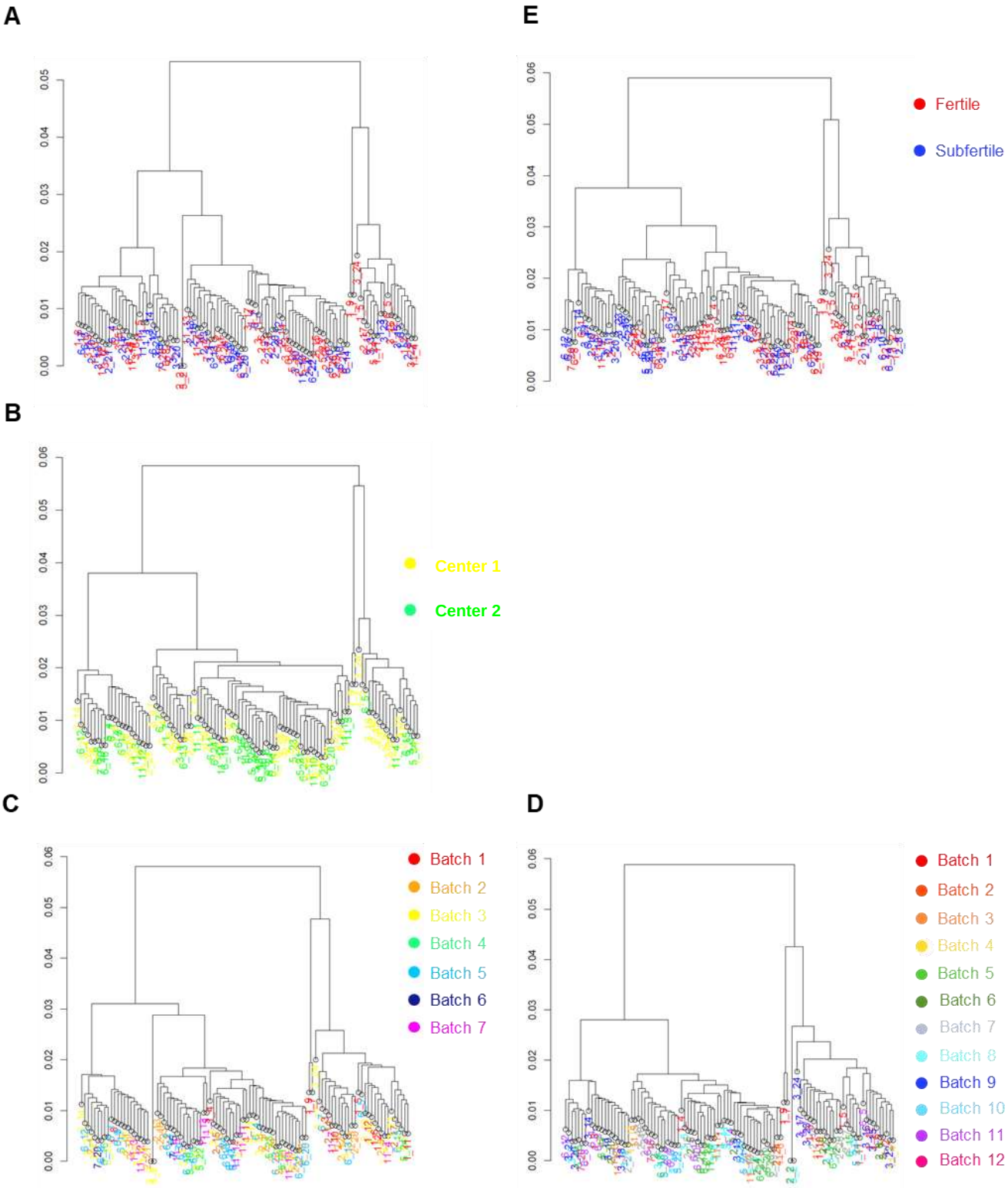
[3]R&D Department, ALLICE, 149 rue de Bercy, 75012, Paris, France.

[4]Université Paris-Saclay, AgroParisTech, INRAE, GABI, 78350 Jouy-en-Josas, France.

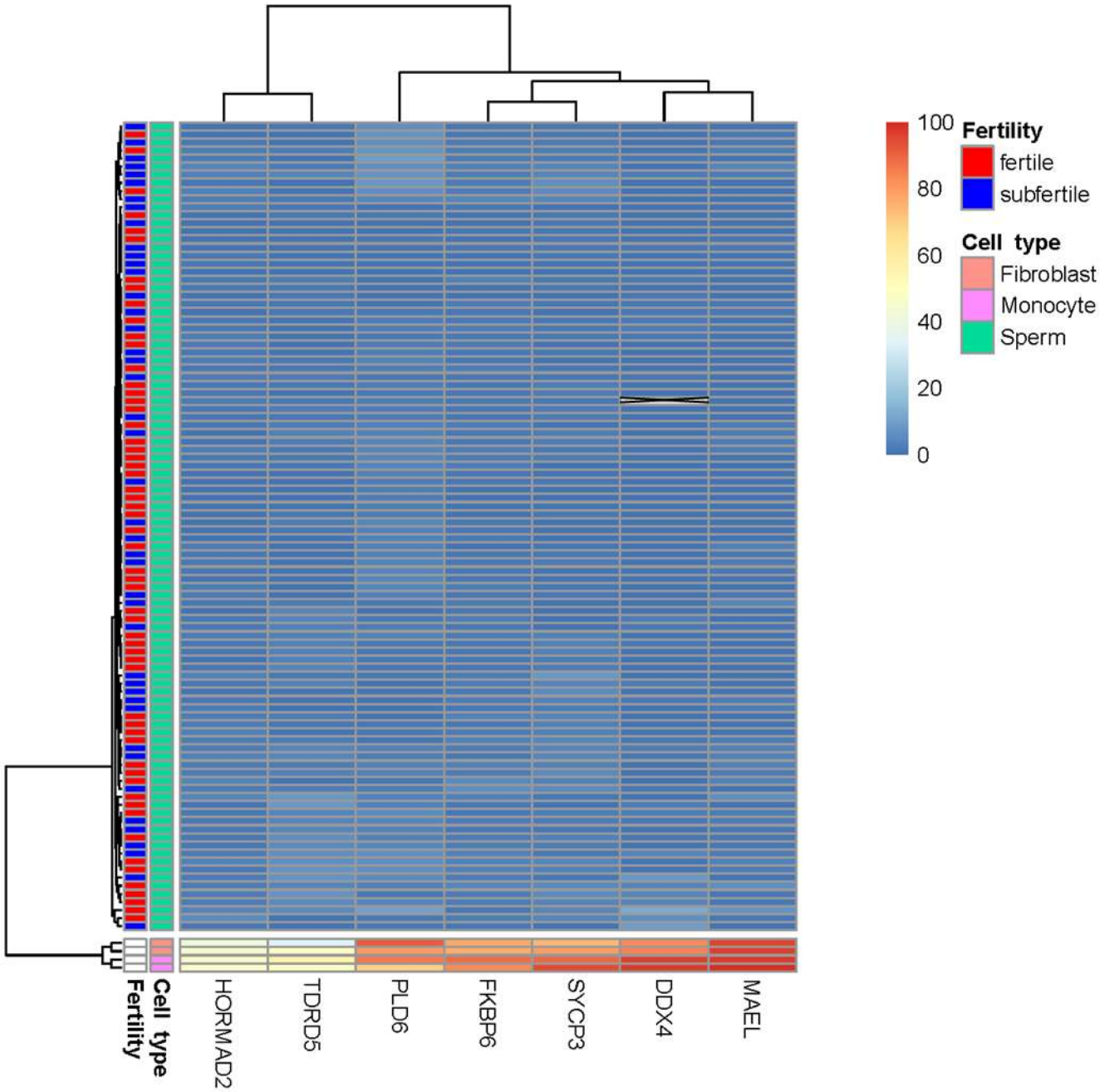
*Corresponding author: helene.kiefer@inrae.fr

Additional results: a larger panel of DMCs with imputed DNA methylation values did not enhance the performance of the predictive model

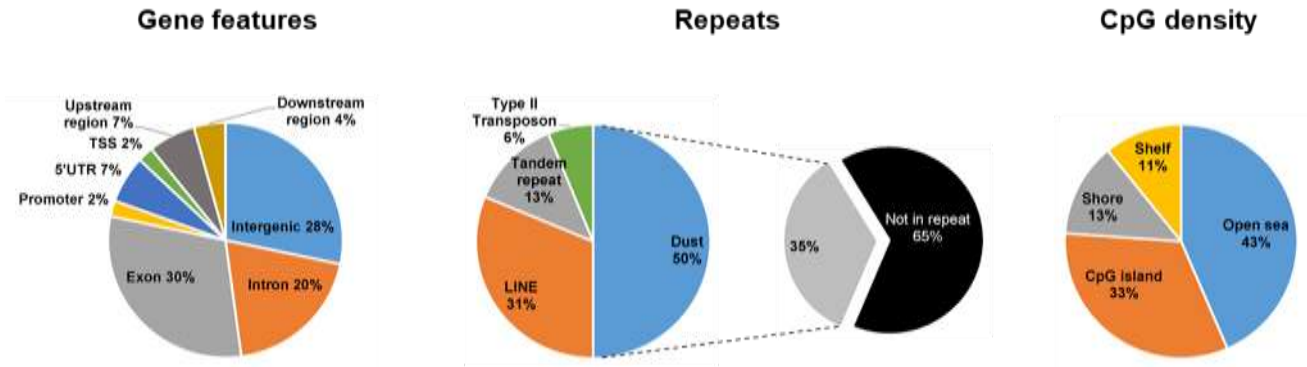
Because of both the large size of the main cohort and the bioinformatics settings, most of the fertility-related DMCs contained missing values (383 out of 490) and were therefore not used to build the model. To verify whether methylation at these DMCs contained additional information that could be used to improve the model, DNA methylation values were imputed to an extended panel of 295 DMCs containing no more than 10% missing values. Principal component analysis run on DMCs without and with imputation did not reveal any significant changes to the percentage of explained variance or in individual factor maps (Supplementary Figure 7). The performance of the new model was also very similar to that obtained using the 107 DMCs without missing values (AUC: 0.83, accuracy: 0.76, sensitivity: 0.83 and specificity: 0.59; averaged values after 50 resamplings of the training and testing sets). These results therefore indicate that DNA methylation at both panels of fertility-related DMCs enabled the prediction of some, but not all, cases of subfertility. The similar results obtained using 107 DMCs with no missing values and 295 DMCs with the imputation of missing values suggest either a certain degree of redundancy among DMCs for the prediction of fertility, or that imputation is not sufficiently precise to provide any additional information. In light of these results, we therefore focused on the model without imputed values.



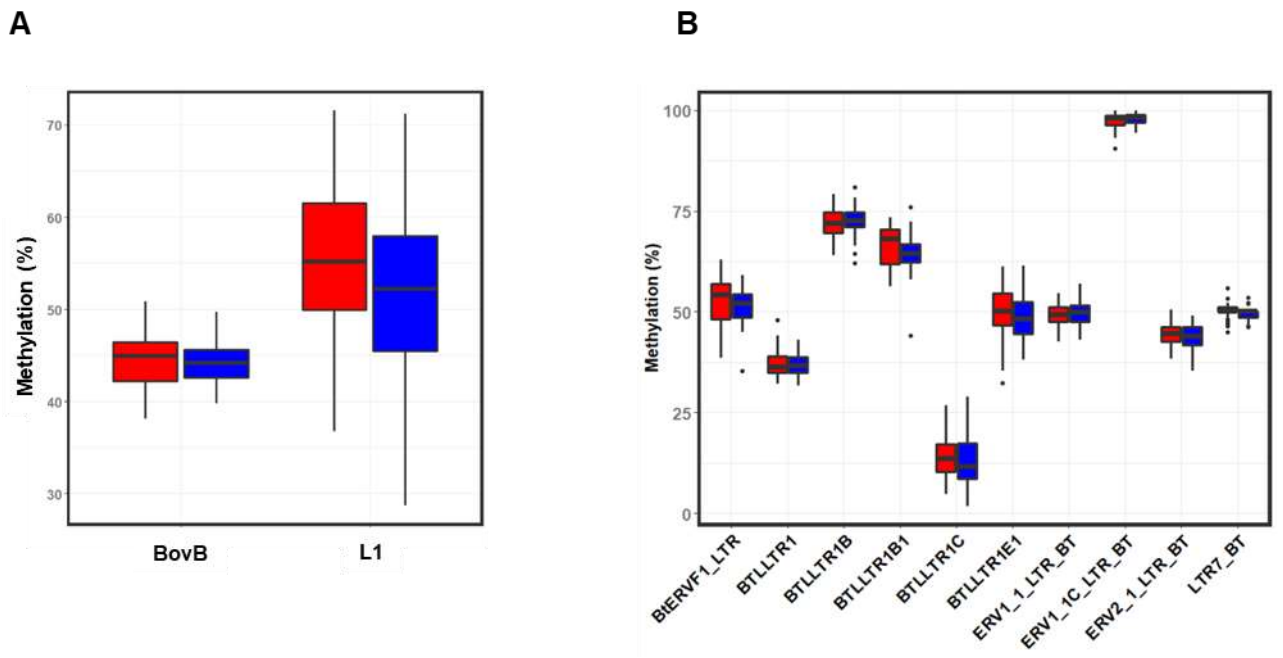
Supplementary Figure 1. Correlation clustering was run on the methylation percentages calculated at CpGs covered by at least 10 reads in at least 22 samples per group in the main cohort, without (**A-D**) or with (**E**) putative sequence variants. **A, E:** samples from the fertile and subfertile bulls are shown in red and blue, respectively. **B:** samples originating from two semen collection centers are shown in yellow and green, respectively. **C:** samples obtained from different semen processing batches are displayed in different colors. **D:** samples obtained from different RRBS library preparation batches are displayed in different colors. Taken together, the results demonstrate that inter-individual variability unrelated to fertility shapes DNA methylation patterns, which are unaffected by confounding effects such as the origins of bulls or technical issues.



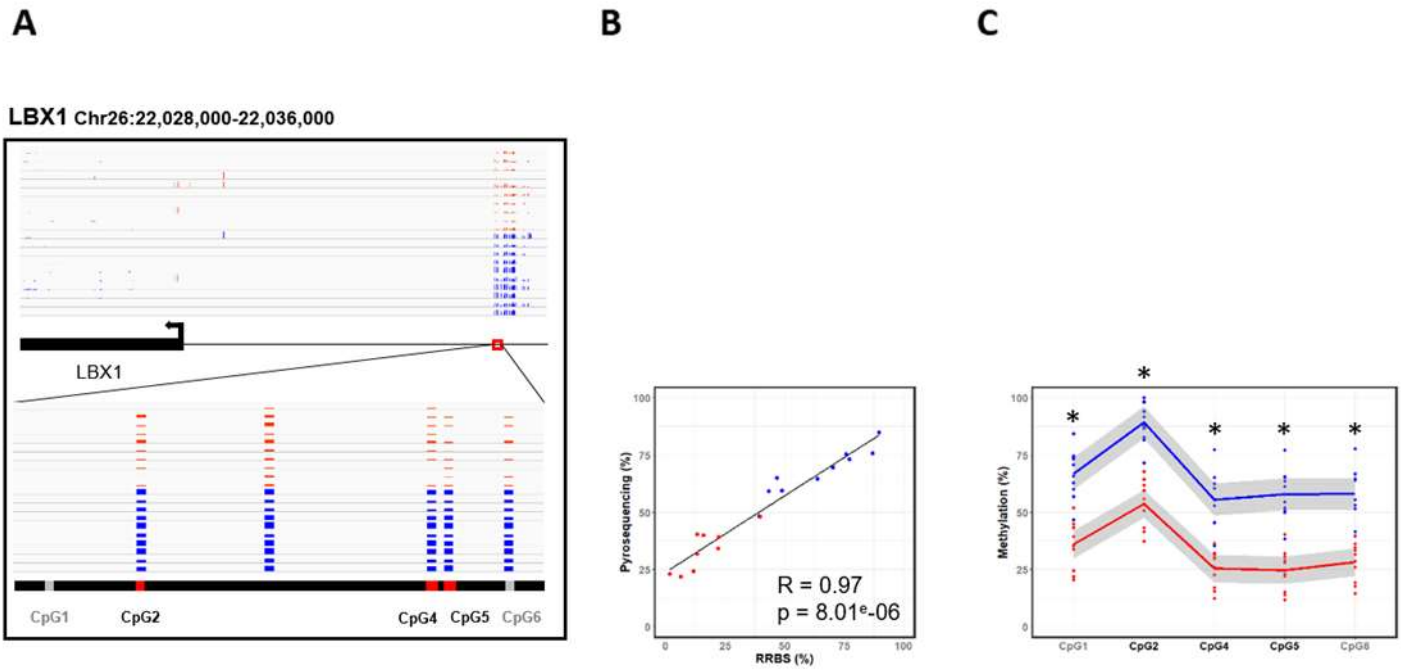
Supplementary Figure 2. Heatmap run on the 100 semen samples in the main cohort and on two types of adult somatic cells (monocytes and fibroblasts). Average DNA methylation values from promoters of genes involved in male gamete generation and at which DNA methylation enables the discrimination of sperm and somatic cells (Perrier et al., 2018) were used to build the heatmap. The black cross indicates that one semen sample was not covered at *DDX4*. Reduced representation bisulfite sequencing data for somatic cells are available under accession GSE102169. The absence of DNA methylation in any of the 100 semen samples confirms that they are not contaminated by somatic cells. Moreover, the 100 semen samples are not clustered according to fertility, demonstrating that residual somatic contamination, if any, does not confound the DNA methylation results related to fertility.



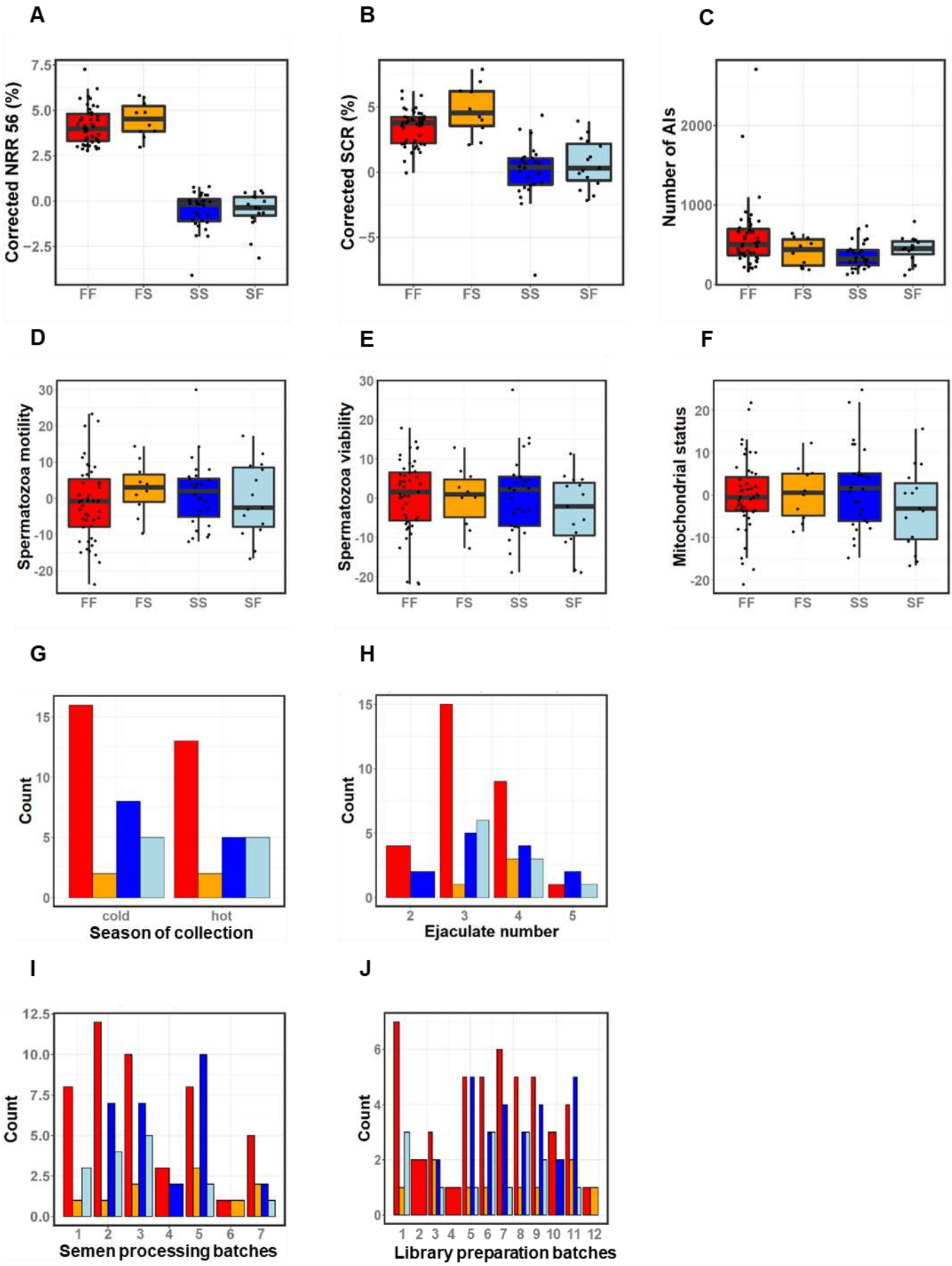
Supplementary Figure 3. The regions differentially methylated between fertile and subfertile bulls were annotated relative to gene features, repeats, and CpG islands, shores and shelves. Consistent with the fact that most of them were hypermethylated in subfertile bulls, the genome features targeted by DMRs are similar to those targeted by the DMCs that were hypermethylated in subfertile bulls (Figure 4A). TSS: transcription start site, UTR: untranslated region, upstream region: from -10 to 0 kb relative to the TSS; downstream region: from 0 to +10 kb relative to the transcription termination site.



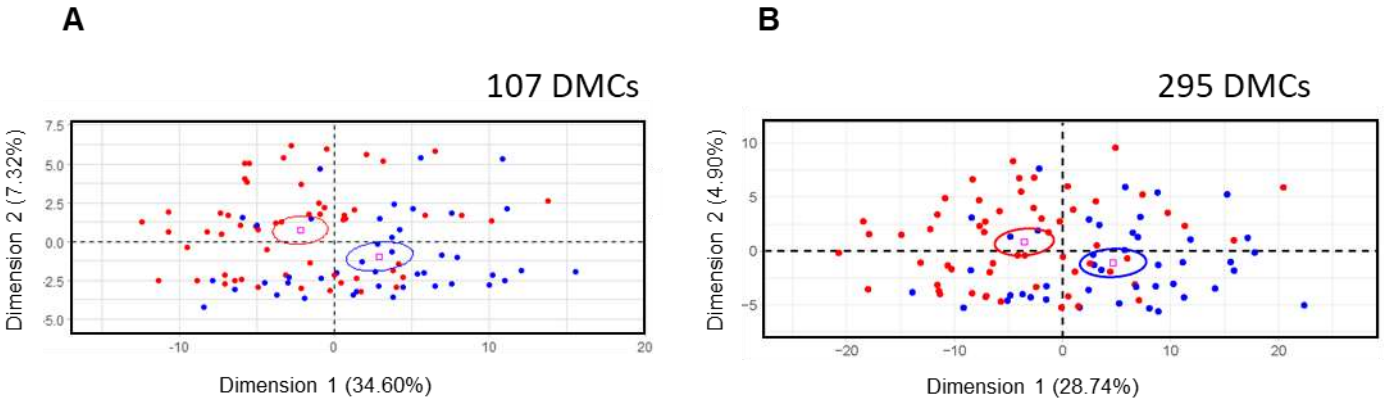
Supplementary Figure 4. RRBS sequences were aligned on a Repbase artificial genome and the average DNA methylation per individual repeat was calculated from the CpGs10 covered in each sample (Supplementary Table 5). Fertility groups were then compared using a Wilcoxon test. **A:** while DNA methylation was identical between fertile and subfertile bulls at LINE BovB, a tendency towards a slight decrease was observed in subfertile bulls at LINE L1 (Wilcoxon test, pvalue=0.072). **B:** the same approach was applied to members of the LTR family, but no difference could be found between fertile and subfertile bulls.



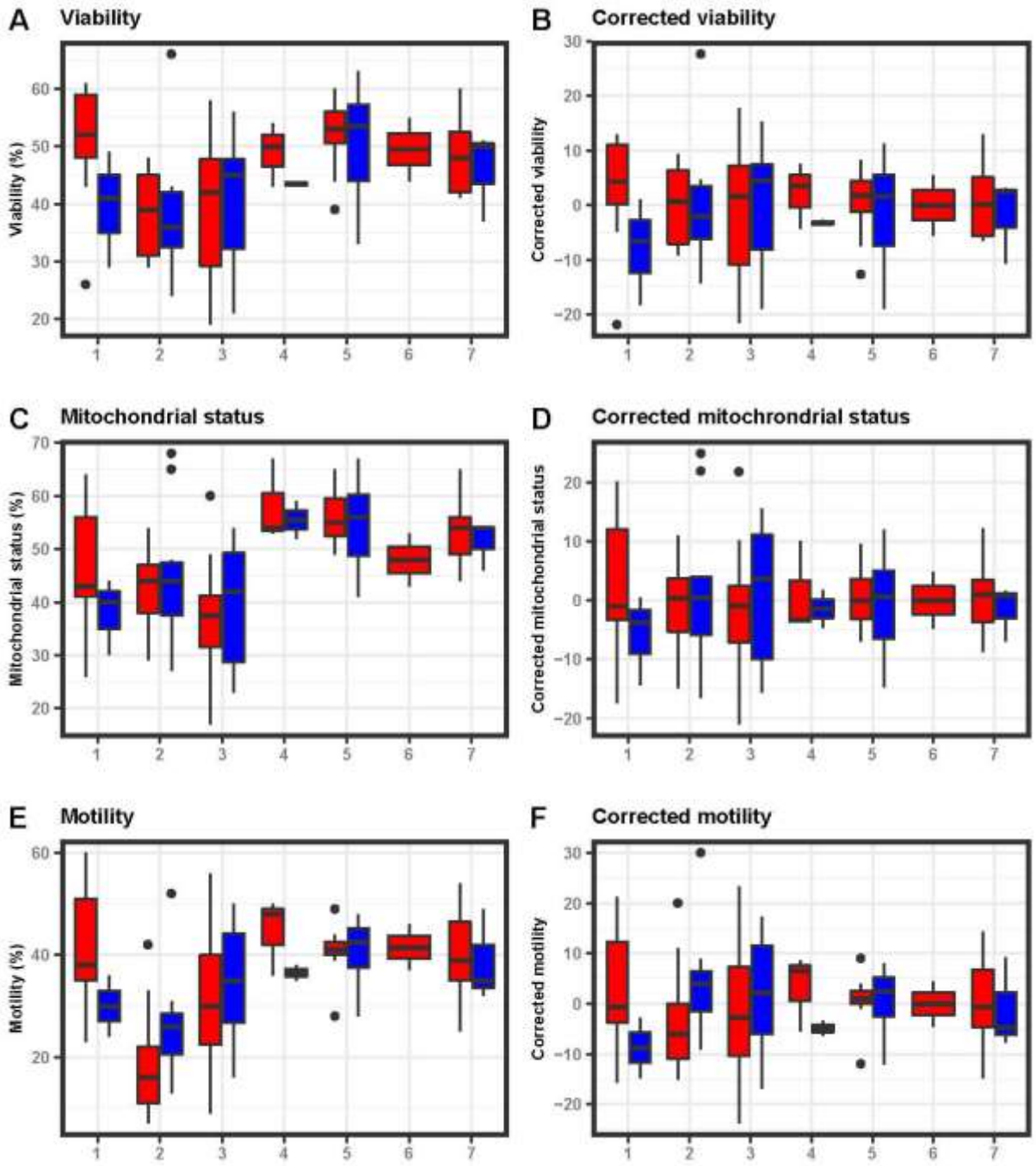
Supplementary Figure 5. A: IGV browser view of the upstream region of the *LBX1* gene. The zoomed view (lower panel) indicates the region targeted by pyrosequencing. The red and blue bar charts represent the methylation percentages at each CpG10 position for fertile (n=10) and subfertile (n=10) bulls, respectively. The CpGs analyzed by pyrosequencing are numbered according to their 5'-3' position along the genome. The CpGs identified as fertility-related DMCs are indicated in black text and red boxes, while non-DMCs are indicated in grey. **B:** The average methylation percentage measured by pyrosequencing (y-axis) was calculated for the three DMCs and plotted against the average methylation percentage measured by RRBS (x-axis). Each dot represents one semen sample from the fertile (in red, n=9) and subfertile (in blue, n=10) groups. The least squares line of best fit and Spearman's R rank correlation coefficient are indicated. **C:** Methylation percentages of individual CpGs assayed by pyrosequencing in fertile (in red, n=10) and subfertile (in blue, n=10) bulls. CpGs are numbered according to **A** and DMCs are highlighted in black. Dots show the methylation levels of individual samples, while the trends per fertility group are indicated by red and blue lines. Asterisks indicate that the methylation percentage measured by pyrosequencing differs significantly between fertility groups for all analyzed CpGs (Wilcoxon test, $p<0.05$).


A
Corrected NRR 56 (%)
7.5
5.0
2.5
0.0
−2.5
FF
FS
SS
SF
B
Corrected SCR (%)
5
0
−5
FF
FS
SS
SF
C
Number of AIs
2000
1000
0
FF
FS
SS
SF
D
Spermatozoa motility
30
20
10
0
-10
-20
FF
FS
SS
SF
E
Spermatozoa viability
30
20
10
0
-10
-20
FF
FS
SS
SF
F
Mitochondrial status
20
10
0
-10
-20
FF
FS
SS
SF
G
Count
15
10
5
0
cold
hot
Season of collection
H
Count
15
10
5
0
2
3
4
5
Ejaculate number
I
Count
12.5
10.0
7.5
5.0
2.5
0.0
1
2
3
4
5
6
7
Semen processing batches
J
Count
6
4
2
0
1 2 3 4 5 6 7 8 9 10 11 12
Library preparation batches

Supplementary Figure 6 (previous page). Within main cohort, correctly predicted bulls (FF: fertile predicted as fertile, in red; SS: subfertile predicted as subfertile, in blue) and misclassified bulls (FS: fertile predicted as subfertile, in orange; SF: subfertile predicted as fertile, in light blue) were compared for field fertility (**A-B**; **A:** corrected NRR 56; non return rate of the cows at 56 days post-insemination; **B:** corrected SCR; sire conception rate), the number of artificial inseminations used to evaluate fertility (**C**); and semen functional parameters (**D-F**; **D:** motility; **E:** viability; **F:** mitochondrial status). No significant differences between misclassified and correctly classified bulls from both fertility classes were observed (Wilcoxon test, adjusted pvalue>0.05). **G-J:** no obvious bias could be observed in the distribution of misclassified and correctly classified bulls regarding the season of semen collection (**G;** cold: from November to April; hot: from May to October; established according to temperature measurements during the years of collection), the number of ejaculates representing each sample (**H**), semen processing batches (**I**) and library preparation batches (**J**), demonstrating the absence of confounding factors in the experimental design.



Supplementary Figure 7. Principal component analysis run on the differentially methylated CpGs identified between fertile (red) and subfertile (blue) bulls belonging to the main cohort. Confidence ellipses are represented. **A:** PCA run on 107 DMCs without missing values. **B:** PCA run on 295 DMCs after imputation of DNA methylation values at DMCs containing no more than 10% missing values. The results were similar in terms of the segregation of fertility groups and the percentages of variance explained by the first two dimensions.



Supplementary Figure 8. Semen functional parameters measured on fertile (red) and subfertile (blue) semen samples before (**A, C, E**) and after correction for the batch effect (**B, D, F**). Due to the large size of the main cohort, semen samples were thawed and analyzed in seven batches of 2-24 samples. The seven batches are shown on the x-axis. The batch effect was no longer visible after correction, making it possible to analyze together the 100 samples whatever the batch during which they have been assayed.

Annexe 4 : Taureaux constituant le dispositif « Fertilité » en race Holstein.

ID	Fertilité	Naissance	Nb éjaculat	Collecte premier éjaculat	Collecte dernier éjaculat	Nb paillette par éjaculat	Sample batch	TNR56	Nb IA
11_9	Fertile	26/05/2014	2	06/11/2015	18/11/2015	5;4	1	4.711398	4123
4_16	Fertile	04/09/2011	9	14/01/2013	16/04/2013	1 pour chaque	2	4.389951	5739
8_6	Fertile	23/03/2011	5	16/08/2012	11/10/2012	2;2;2;2;1	3	4.065883	3410
4_21	Fertile	02/05/2012	6	24/08/2012	26/10/2012	2;1;1;2;2;1	2	3.667029	9056
8_5	Fertile	19/02/2011	5	05/07/2012	27/08/2012	2;2;2;2;1	3	3.294712	9250
4_5	Fertile	06/12/2011	9	23/04/2013	19/07/2013	1 pour chaque	2	3.272248	10636
4_12	Fertile	31/07/2011	5	04/12/2012	26/12/2012	2;2;2;1;1	2	3.267768	5702
10_6	Fertile	19/04/2014	2	07/09/2015	23/10/2015	2;7	4	3.202356	22765
4_10	Fertile	01/07/2011	9	03/12/2012	01/02/2013	1 pour chaque	2	3.192631	19847
8_11	Fertile	26/11/2014	1	06/06/2016	06/06/2016	8;	3	3.180203	10205
4_23	Fertile	22/08/2011	8	04/12/2012	28/03/2013	2;1;1;1;1;1;1;1	2	3.178289	23812
8_15	Fertile	29/09/2011	5	08/02/2013	19/03/2013	2;2;2;2;2	3	3.168653	6695
8_7	Fertile	27/05/2011	6	22/11/2012	31/12/2012	1;1;1;1;2;3	3	2.970556	20304
10_16	Fertile	09/01/2014	1	26/05/2015	26/05/2015	9;	4	2.782284	2547
8_17	Fertile	16/01/2012	1	22/08/2013	22/08/2013	9;	3	2.674632	3601
8_4	Fertile	16/02/2011	7	19/07/2012	27/09/2012	2;1;1;1;1;2;1	3	2.660827	6218
4_14	Fertile	06/01/2012	2	24/04/2013	24/06/2013	5;4	2	2.635039	11804
4_18	Fertile	24/03/2011	3	06/09/2012	17/09/2012	3;3;3	2	2.573714	976
8_24	Fertile	14/05/2014	2	07/12/2015	24/12/2015	4;5	3	2.562471	4668
4_2	Fertile	30/04/2011	9	27/08/2012	23/11/2012	1 pour chaque	2	2.508123	6182
8_8	Subertile	30/08/2011	5	22/01/2013	28/02/2013	2;2;2;2;2	3	0.690702	2487
4_4	Subertile	11/12/2011	6	27/05/2013	19/06/2013	2;2;1;2;1;1	2	0.523005	7059
8_23	Subertile	28/06/2011	3	12/11/2012	14/01/2013	3;3;3	3	0.506559	1387
4_19	Subertile	30/07/2011	7	13/11/2012	22/02/2013	1;1;1;1;2;2;1	2	0.400932	11023
4_22	Subertile	03/04/2012	9	02/08/2013	24/09/2013	1 pour chaque	2	0.308154	3590
8_22	Subertile	14/11/2014	2	03/06/2016	17/06/2016	4;5	3	0.30726	14481

4_1	Subertile	29/06/2011	3	05/12/2012	26/12/2012	3;3;3	2	-0.12369	6407
8_2	Subertile	03/03/2011	4	14/08/2012	31/08/2012	2;3;2;2	3	-0.34289	3157
4_15	Subertile	18/03/2012	3	20/08/2013	20/09/2013	3;3;3	2	-0.38999	3350
4_11	Subertile	12/05/2011	9	21/09/2012	30/11/2012	1 pour chaque	2	-0.63234	7122
8_18	Subertile	09/05/2011	4	06/09/2012	20/09/2012	2;2;3;1	3	-0.7824	1947
8_3	Subertile	30/06/2011	3	14/01/2013	24/01/2013	3;3;3	3	-0.81142	2555
8_9	Subertile	22/04/2012	2	13/09/2013	17/09/2013	5;4	3	-0.82626	3507
10_19	Subertile	27/04/2012	1	01/10/2013	01/10/2013	NA	4	-0.95222	816
8_19	Subertile	04/03/2011	3	24/07/2012	14/09/2012	3;3;3	3	-0.99035	2018
8_12	Subertile	29/04/2011	1	16/10/2012	16/10/2012	9;	3	-1.20186	809
10_20	Subertile	15/11/2014	2	15/04/2016	17/05/2016	5;4	4	-1.34726	4087
10_17	Subertile	17/12/2014	3	01/06/2016	18/07/2016	3;3;3	4	-1.57348	4860

Annexe 5 : Qualité des séquences et des alignements de la cohorte « Fertilité » en race Holstein

	Fertile (n = 20)	Subfertile (n = 18)
Paramètres séquençage		
Nombre de paires de séquences (million)	34.6 +- 5.6	34 +- 4.6
Qualité des séquences (score Phred)	38.4 +- 0.2	38.1 +- 0.3
Conversion Bisulfite (%)	98.76 +- 0.3	98.76 +- 0.3
Alignement ARS-UCD2.1		
Alignement unique (%)	37.04 +- 1.5	36.93 +- 2.3
Nombre de CpG couverts (million)	3.3 +- 0.08	3.2 +- 0.08
Couverture moyenne par CpG	25.6 +- 4	25.4 +- 3.4
Pourcentage de CpG10	63.5 +- 3.6	63.1 +- 3.4
Méthylation moyenne CpG10 (%)	47.92 +- 0.8	48 +- 0.9
Pourcentage de CpG10 hypométhylées	49.06 +- 0.9	49.01 +- 0.9
Pourcentage de CpG10 intermédiaires	6.07 +- 0.3	6.1 +- 0.23
Pourcentage de CpG10 hyperméthylées	44.88 +- 0.8	44.9 +- 0.9

Annexe 6 Données supplémentaires de l'article 2

Omics data integration for bull fertility prediction

Valentin Costes[1,2,3,4], Eli Sellem[1,2,3], Sylvain Marthey[4,5], Chris Hoze[3,4], Aurélie Bonnet[1,2,3], Laurent Schibler[3], Hélène Kiefer[1,2] and Florence Jaffrezic[4*]

[1]Université Paris-Saclay, UVSQ, INRAE, BREED, 78350 Jouy-en-Josas, France.

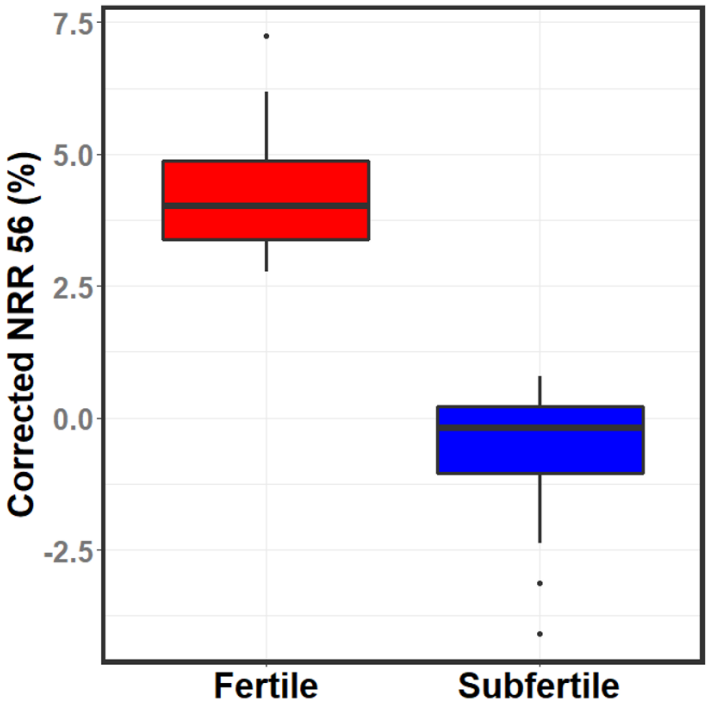
[2]Ecole Nationale Vétérinaire d'Alfort, BREED, 94700, Maisons-Alfort, France.

[3]R&D Department, ALLICE, 149 rue de Bercy, 75012, Paris, France.

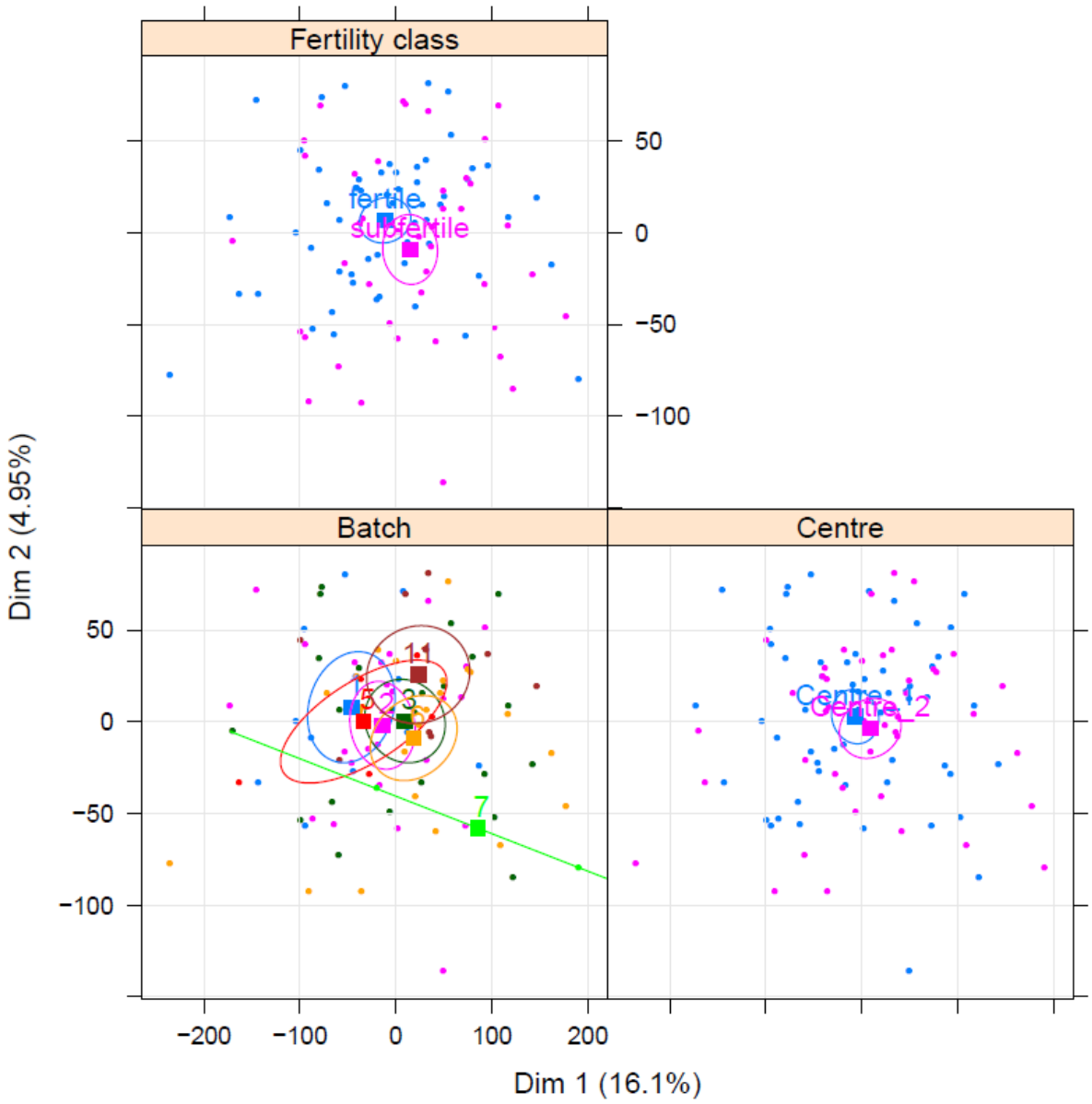
[4]Université Paris-Saclay, AgroParisTech, INRAE, GABI, 78350 Jouy-en-Josas, France.

[5]INRAE, MaIAGE, Université Paris-Saclay, 78350 Jouy-en-Josas, France.

*Corresponding author: florence.jaffrezic@inrae.fr

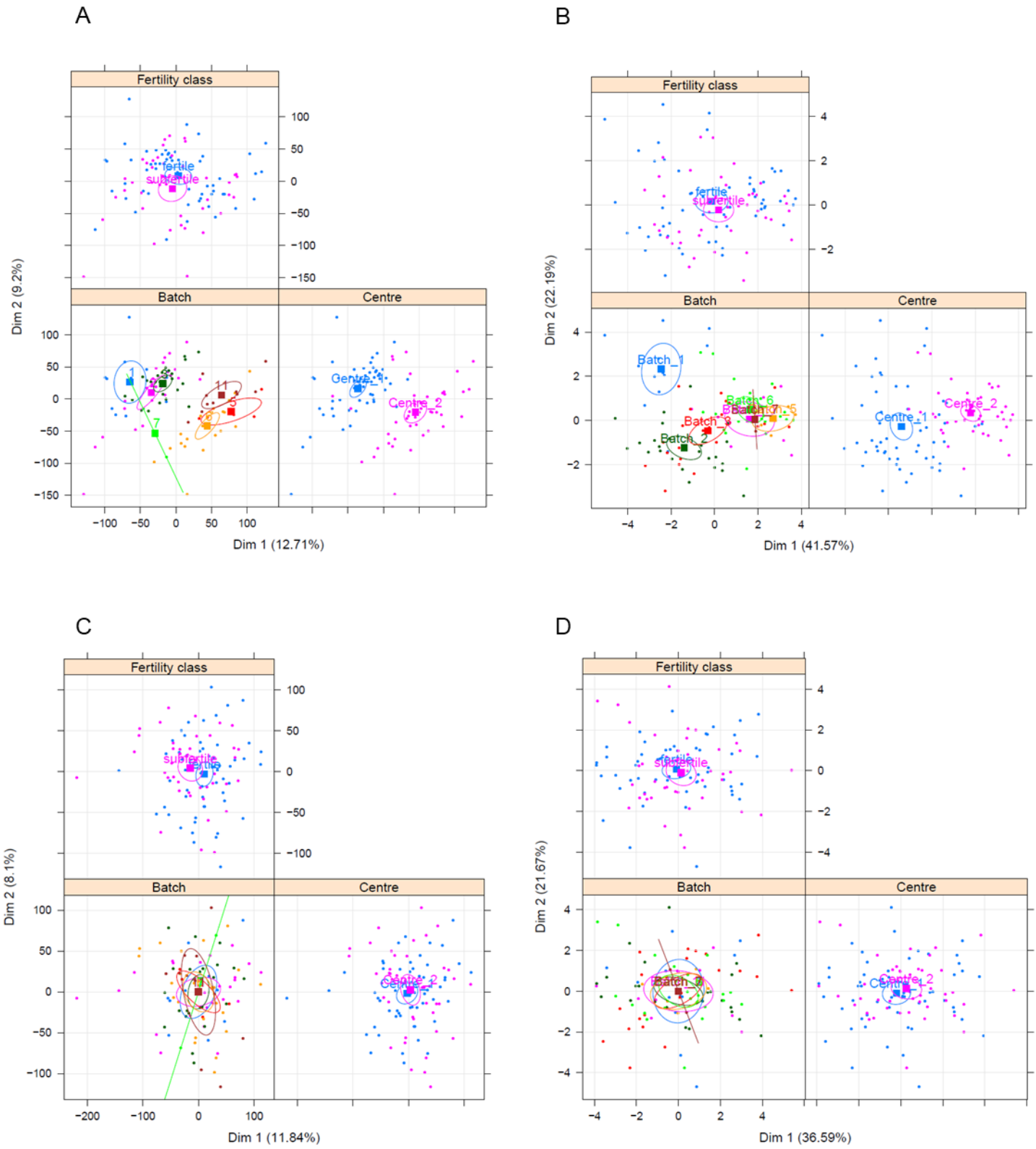


Supplementary Figure 1: Corrected non return rates at 56 days (NNR 56) in fertile (red) and subfertile (blue) bulls.



Supplementary Figure 2: A PCA was run on the 40,000 CpGs remaining after the pre-filtering step and the two first dimensions are represented. The same PCA was illustrated by the fertility class (upper panel), the batch effect (lower left panel) and the semen collection centre (lower right panel). The methylation values at these CpGs were not influenced by the batches and the semen collection centres, and were therefore not corrected for these effects.

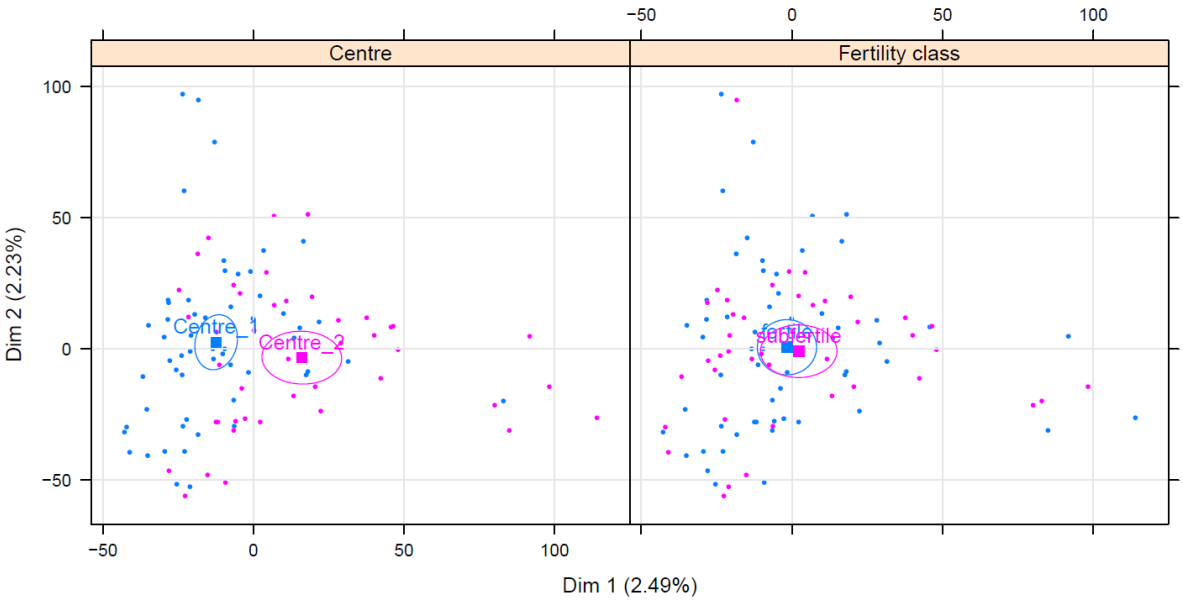
44


A
Fertility class
fertile
subfertile
Batch
Centre
Centre_1
Centre_2
Dim 1 (12.71%)
Dim 2 (9.2%)
B
Batch_1
Batch_2
Batch_6
Batch_7
Dim 1 (41.57%)
Dim 2 (22.19%)
C
Dim 1 (11.84%)
Dim 2 (8.1%)
D
Batch_9
Dim 1 (36.59%)
Dim 2 (21.67%)

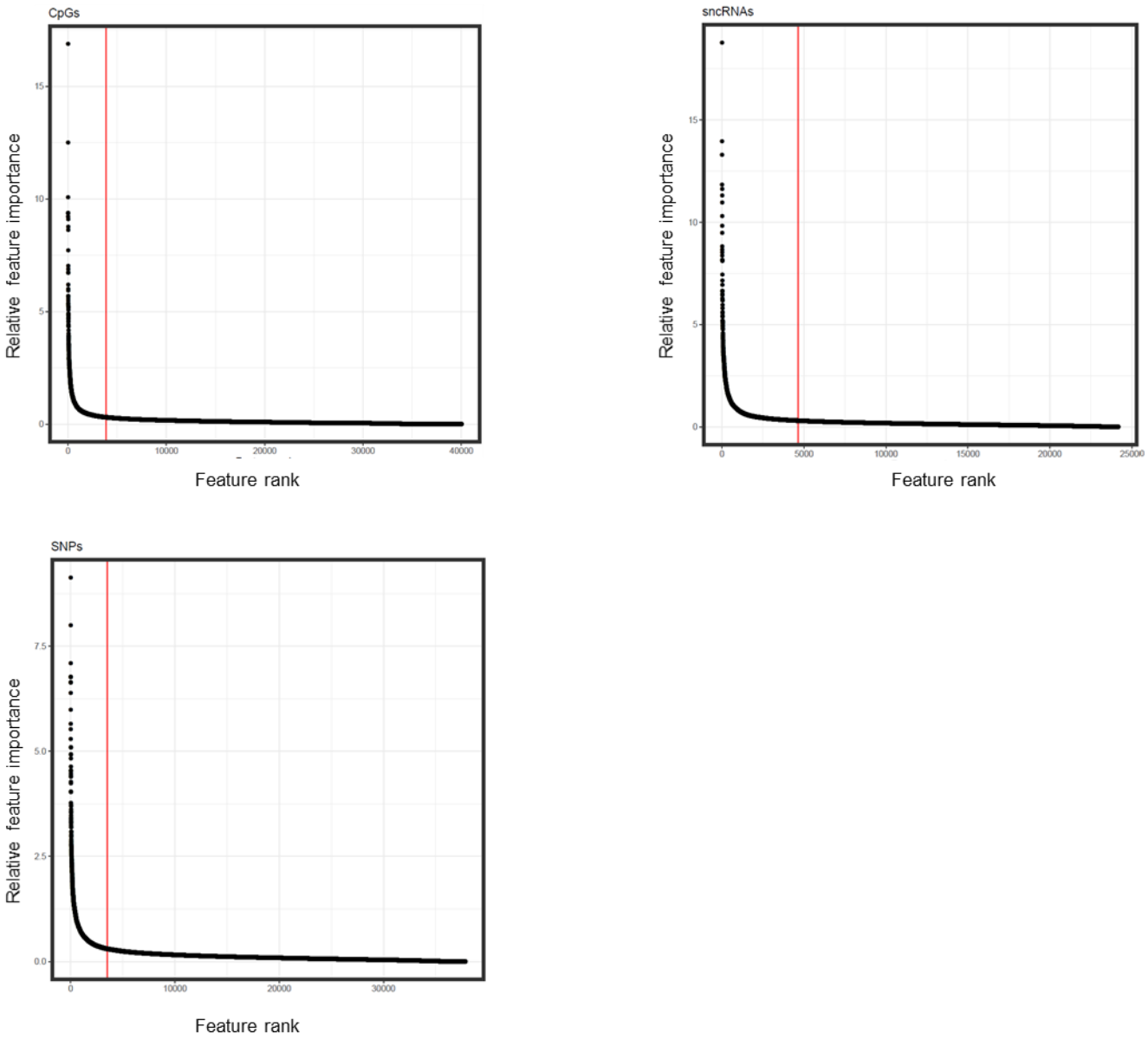
45

46

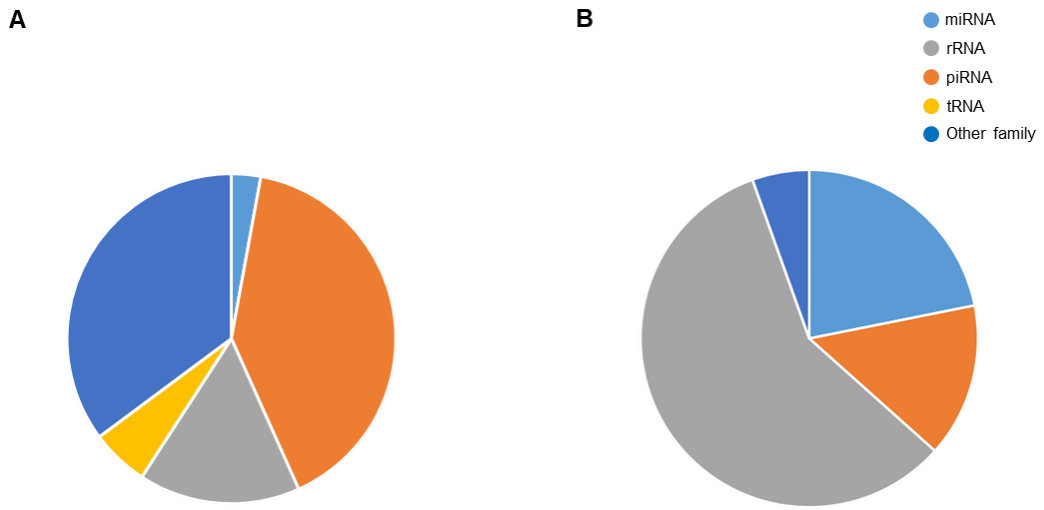
Supplementary Figure 3 (previous page): A PCA was run on 24,172 sncRNAs (A) and 11 semen parameters (SPs) (B) without correction. The two first dimensions are represented and three different factors were used as illustrating variables: the fertility class, the experimental batch and the semen collection centre. The experimental batch and the semen collection centre have both a huge effect on sncRNAs and SPs. After correction for the batch effect, a PCA was run on the scnRNAs (C) and SPs (D) corrected data. The corrected data are no longer biased according to the batch nor the centre.



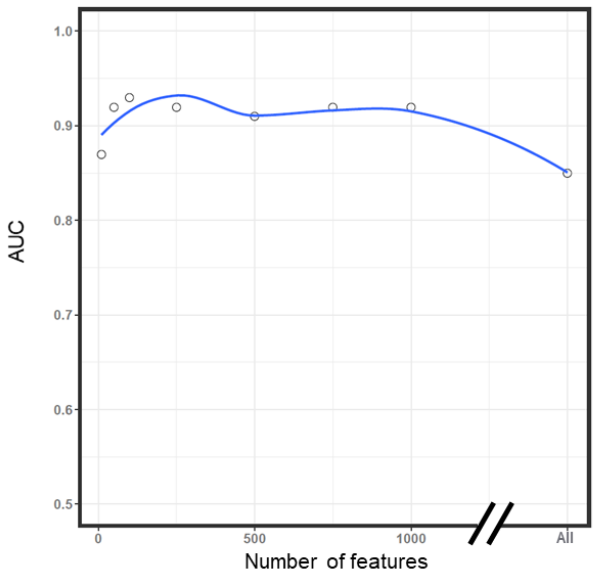
Supplementary Figure 4: A PCA was run on the 38,853 SNPs remaining after the pre-filtering step and the two first dimensions are represented. The same PCA was illustrated by the semen collection centre (left panel) and the fertility class (right panel).



Supplementary Figure 5: For each type of data, the relative feature importance of the model was plotted against the feature rank. The red vertical line indicates the threshold fixed for the pre-selection by Random Forest, where each feature on the left of this curve was selected.



Supplementary Figure 6: Pie charts showing the proportion of the different classes of sncRNAs in the background (A) and in clusters 1 and 2 that were highlighted by the MFA (Figure 2).

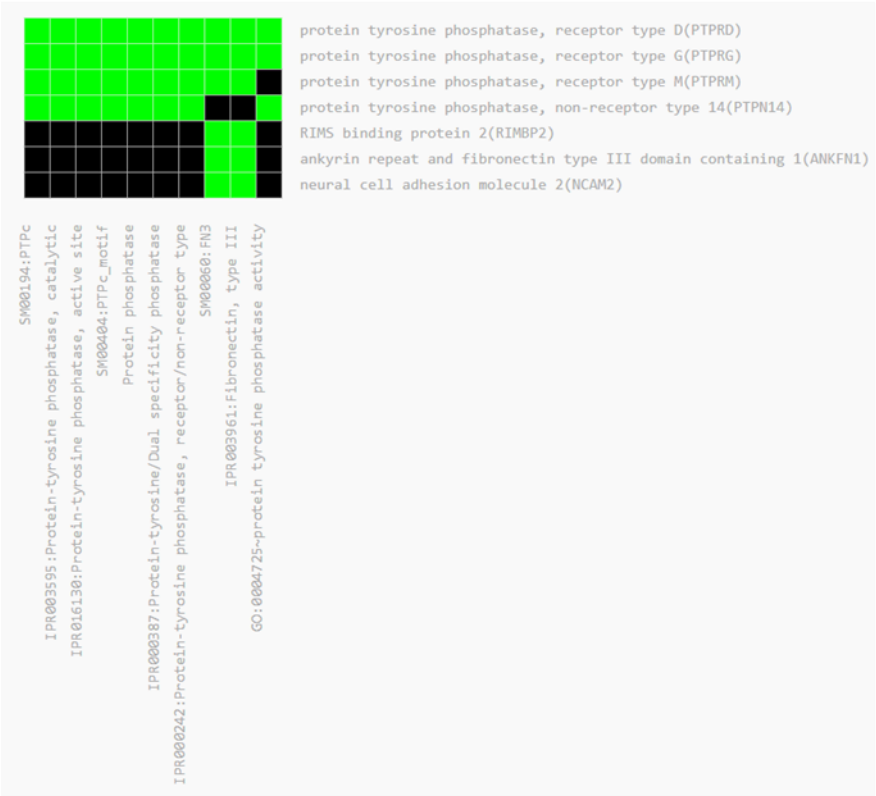


	Number of optimal features	AUC
Cforest	100	0.92

Supplementary Figure 7: One model was constructed using the cforest method with the 12,006 features and features were classified depending on their importance. Then, models were constructed with the top 1000, 750, 500, 250, 100, 50 and 10 features. Using this information, the figure on the left hand indicates the AUC on the y-axis and the number of features used during model construction on the x-axis. Each dot represents the actual AUC value obtained for each model. A tendency curve was drawn using the geom_smooth function of the ggplot2 package with default parameters. The table on the right hand shows the optimal number of features and the associated AUC value obtained for the cforest method, based on the actual AUC values and not on the tendency curve.

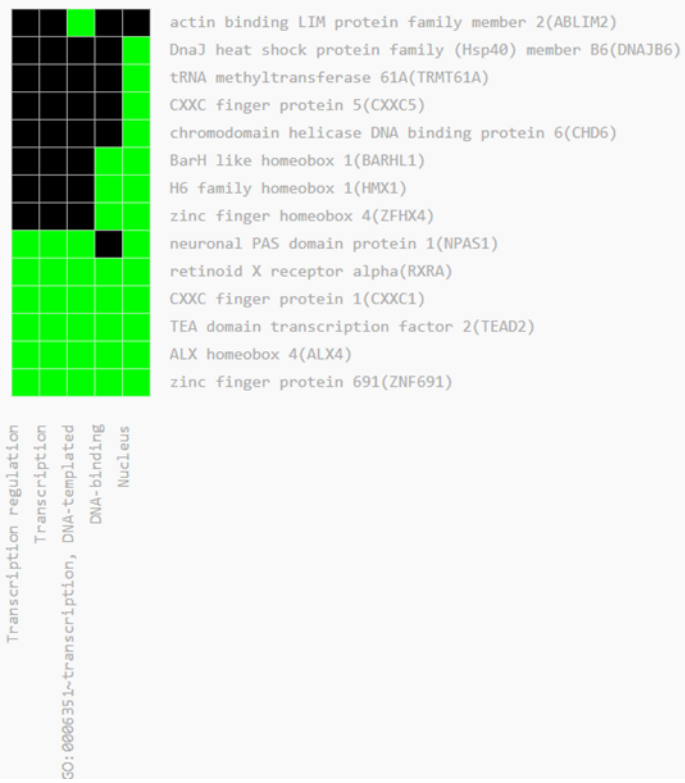
A

EASE score 1.68

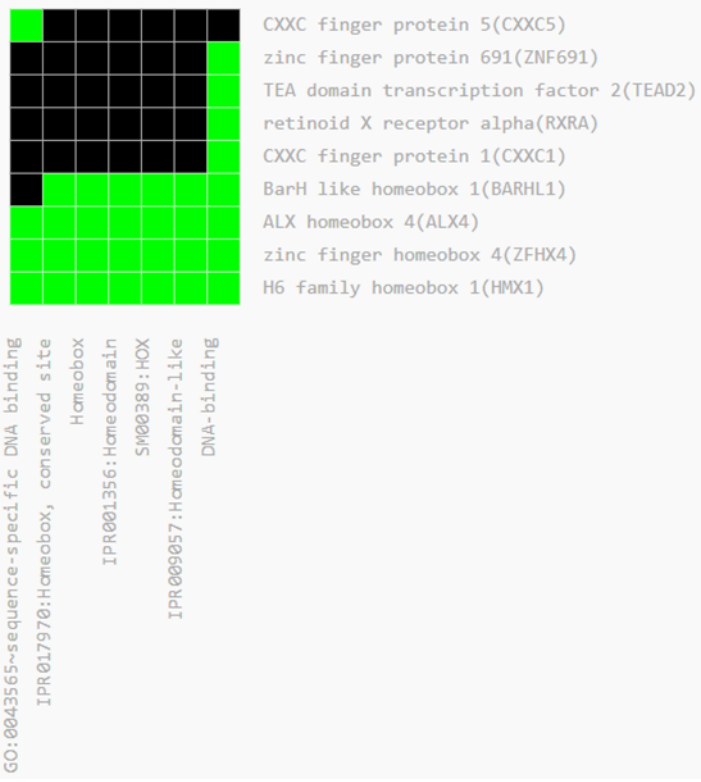


B

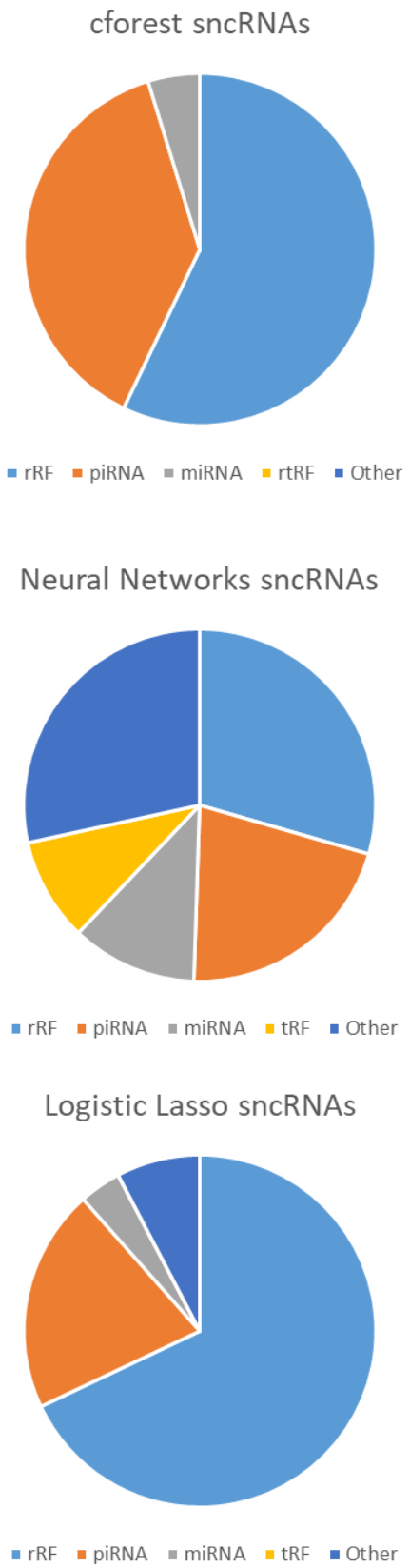
EASE score 1.45



EASE score 1.45



Supplementary Figure 8 (previous page): The genes containing SNP and CpG features selected by Neural Networks (A) and Logistic Lasso (B) were submitted to an enrichment analysis using DAVID. Three clusters of terms were significantly enriched (EASE score above 1.3).



Supplementary Figure 9: Distribution of the different sncRNA families among the sncRNA features identified by each method individually.

Annexe 7 : Article : « The epigenome of male germ cells and the programming of phenotypes in cattle ».

Feature Article

The epigenome of male germ cells and the programming of phenotypes in cattle

Hélène Kiefer,[†,‡] Eli Sellem,[||] Amélie Bonnet-Garnier,[†,‡] Maëlle Pannetier,[†,‡] Valentin Costes,[†,‡,||] Laurent Schibler,[||] and Hélène Jammes[†,‡]

[†]Université Paris-Saclay, UVSQ, INRAE, BREED, 78350, Jouy-en-Josas, France
[‡]Ecole Nationale Vétérinaire d'Alfort, BREED, 94700, Maisons-Alfort, France
[||]R&D Department, ALLICE, 149 rue de Bercy, 75012, Paris, France

Downloaded from https://academic.oup.com/af/article/11/6/28/6469044 by INRA Avignon user on 21 September 2022

Implications

- Bull semen is a commercial product widely used for artificial insemination.
- During the differentiation of male germ cells into spermatozoa, there are several windows of epigenome sensitivity to environmental factors.
- The epigenome of bull sperm exhibits both conserved features and interindividual variations, some of which are associated with fertility.
- The paternal epigenome contributes to embryo development and to programming the phenotype of offspring.

Key words: cattle, DNA methylation, embryo development, epigenetics, male germ cells, small noncoding RNAs

Introduction

DNA methylation plays an important role in the structure and stability of the genome when associated with heterochromatin and repetitive elements such as transposable elements (**TE**s; mobile elements in the genome of viral origin that have accumulated during evolution) and satellites (sequences located near or at the centromeric regions of chromosomes). DNA methylation also regulates transcription and is dynamic as a function of cell type, developmental stage, the animal's physiology, or the environment. Finally, DNA methylation is involved in the genomic imprinting of genes expressed in a mono-allelic manner depending on the parental origin of the allele, a phenomenon that is essential for harmonious growth of the fetus. DNA methylation is catalyzed by DNA methyltransferases (**DNMT**s) that use S-adenosylmethionine as a methyl donor (produced from dietary folic acid). DNMT3A and DNMT3B enzymes are involved in de novo DNA methylation, while DNMT1, which recognizes the hemi-methylated cytosine-phosphate-guanine (**CpG**) sites resulting from DNA replication, ensures the maintenance and propagation of methylation patterns through cell division. The erasure of DNA methylation can result not only from a lack of DNMT1 activity but also from the conversion of 5-methylcytosines (**5meC**s) by ten-eleven translocation (**TET**) enzymes into oxidized derivatives such as 5-hydroxymethylcytosine (**5hmC**), which are then diluted during replication or replaced by unmethylated cytosines by the DNA repair machinery. Genomic DNA is wrapped around octamers of histones to form the nucleosome. Histones contain N-terminal tails targeted by different types of modifications, such as acetylation and methylation. These modifications affect different amino acids, producing dozens of posttranslational variants with different functional roles. The addition or removal of these modifications is highly flexible processes that directly affect the accessibility of genomic DNA to the transcription machinery and hence the activation or repression of gene expression. The combinatorial nature of the different histone marks, together with DNA methylation and the presence of certain transcription factors or RNA polymerase, define specific chromatin states associated with specific transcriptional states. These chromatin states are transmitted to daughter cells, thus ensuring the continuity of cell identity through mitosis. Finally, small noncoding RNAs (**sncRNA**s) play an important role in posttranscriptional regulations, and their expression is tightly regulated in a cell type-specific manner.

Mature spermatozoa are transcriptionally inactive and represent the ultimate form of male germ cell (**GC**) differentiation. Their fate is to survive outside the organism and contribute to a new individual after fertilization of an oocyte. In support of these functions, the epigenome of spermatozoa is unique (Carrell, 2012). Depending on the species, 85% to 99% of the histones are replaced by protamines, arginine-rich proteins that form toroid-shaped structures with DNA. This replacement enables a higher level of chromatin compaction, which contributes to reducing nuclear volume and helps to protect the paternal genetic heritage against oxidation during migration through the epididymis and female genital tract (Champroux et al., 2016). Furthermore, in addition to microRNAs (**miRNA**s) and small interfering RNAs (**siRNA**s) that are also found abundantly in

© Kiefer, Sellem, Bonnet-Garnier, Pannetier, Costes, Schibler, Jammes

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
https://doi.org/10.1093/af/vfab062

somatic cells, the germline is enriched in P-element induced wimpy testis (**PIWI**)-interacting RNAs (**piRNA**s).

The sperm-specific epigenome is acquired during the differentiation of male GCs into mature spermatozoa, a process that starts at the embryonic stage and is only achieved after puberty has been reached and during each cycle of spermatogenesis (Carrell, 2012; Champroux et al., 2016). In cattle and especially dairy breeds, where bull semen is widely used for artificial insemination (**AI**), several selection and breeding practices may interfere with proper establishment of the sperm epigenome (Figure 1). The selection of AI bulls relies on their genetic merit, and they are usually obtained from the breeding of high breeding value sires and high-producing dairy cows. These cows are more likely to experience a negative energy balance in the event of concurrent lactation and gestation, which may lead to an unfavorable in utero environment for the developing fetus (Wu and Sirard, 2020). Otherwise, practices to reduce the generation interval and accelerate genetic gain, such as hormonal treatments of the mothers, embryo technologies, or the hastened growth and puberty of male calves, may have a long-term impact on the sperm epigenome (Rivera, 2019). Finally, bull semen is extensively processed before its use for AI, which, according to data obtained in other species, may affect the chromatin structure (Aurich et al., 2016).

Because bull semen has a widespread diffusion potential, with dozens of offspring potentially being generated per batch, it is important to understand the impact of these practices on the epigenetic landscape of spermatozoa and the degree to which variations in the epigenome might affect fertility and the phenotype of offspring. The goal of this short review is, therefore, to provide an overview of recent knowledge regarding the epigenome of male GCs and its potential role in the programming of phenotypes, with particular emphasis on cattle and in light of the knowledge accumulated in other species.

Epigenetic Reprogramming of Male GCs and Windows of Sensitivity to Environmental Factors

Establishment of the male germline requires three successive stages: 1) specification of primordial GCs (**PGC**s) from the embryonic epiblast, 2) migration and colonization of the genital ridges that will form the testes, and 3) differentiation into male GCs (pro-spermatogonia or gonocytes), which stop proliferating and enter quiescence. After birth, the male GCs resume mitosis and progressively migrate from the center to the basement of seminiferous cords. In parallel with these processes, the pool of spermatogonial stem cells (**SSC**s), from which spermatogenesis is sustained over a lifetime, is gradually established from male GCs. The whole process of spermatogenesis only becomes effective after puberty and comprises a mitotic phase (spermatogonia), a meiotic phase (spermatocytes), and spermiogenesis (spermatids). During this last step, dramatic morphological changes occur that convert round and transcriptionally active spermatids into spermatozoa harboring a flagellum, a head, a tightly compacted nucleus, an acrosome, and almost no cytoplasm. Spermatozoa are then released into the lumen of the seminiferous tubules and transit through epididyma where they acquire motility and complete their maturation until fertilization (Wrobel, 2000; Staub and Johnson, 2018).

These differentiation and maturation processes are based on a specific transcriptional program orchestrated by extensive epigenetic reprogramming (Figure 2). Most knowledge concerning the reprogramming of DNA methylation has been acquired in mice and, to a lesser extent, in humans. The DNA methylation pattern that characterizes the epiblast is thus erased throughout the genome when PGCs colonize the gonad. DNA methylation erasure involves mechanisms that are both passive (through cell division and the absence of maintenance activity) and active (through the generation of 5hmC). This erasure is not total: some genomic regions retain methylation to

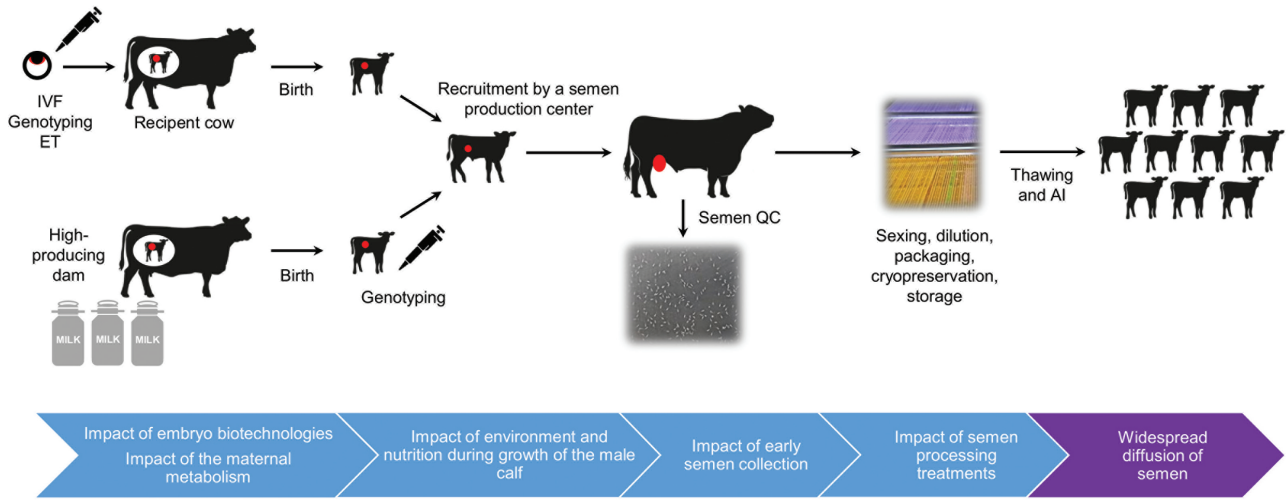


Figure 1. Breeding, selection, and semen processing practices in the cattle AI industry. The different steps that may impact the sperm epigenome of bulls are highlighted. The developing germline is shown in red. ET, embryo transfer; IVF, in vitro fertilization; QC, quality control.

Downloaded from https://academic.oup.com/af/article/11/6/28/6469044 by INRA Avignon user on 21 September 2022

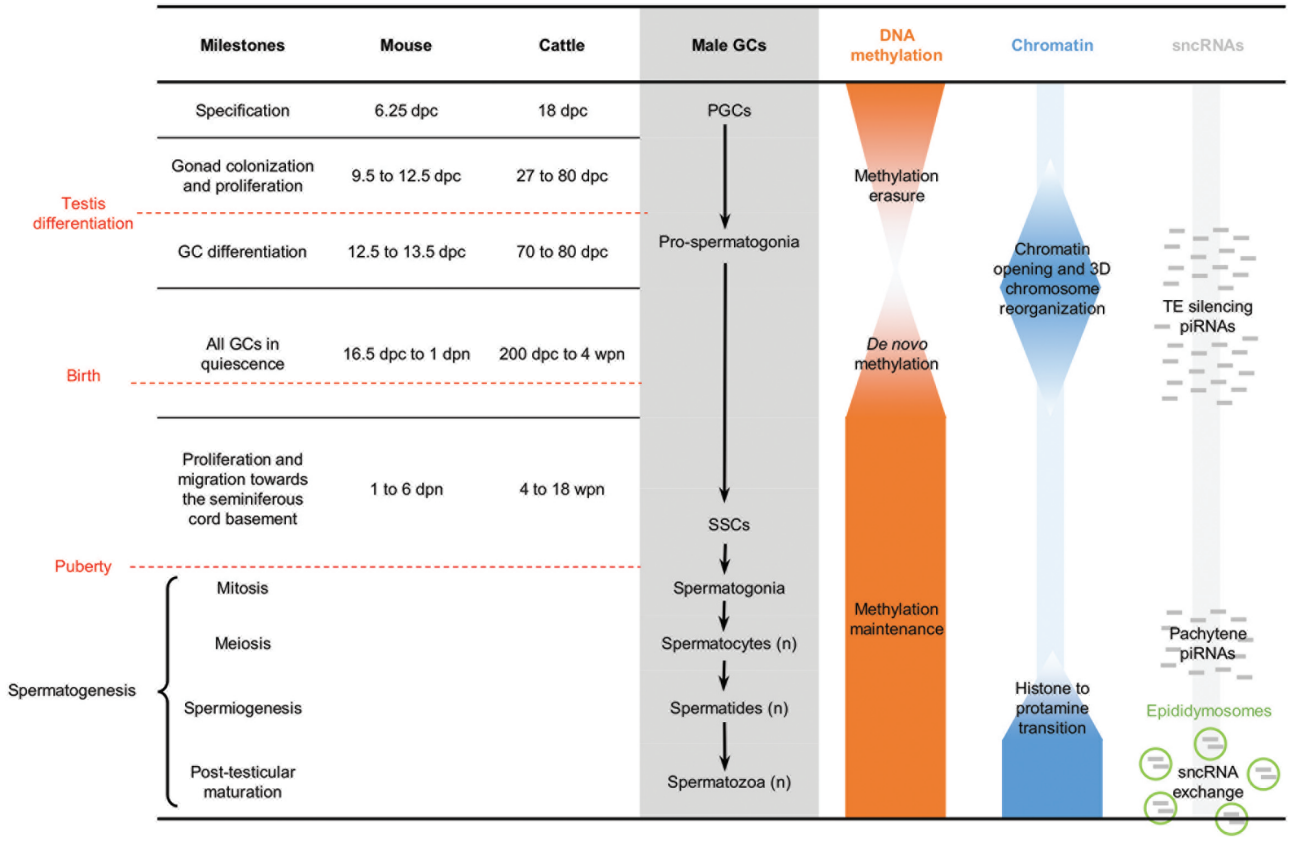


Figure 2. Epigenetic reprogramming during the differentiation of male GCs in mice. The timing of the different milestones in cattle has been established using unpublished data from our lab and from the study of Wrobel (2000). The timing of epigenetic reprogramming in bovine fetuses is unknown. dpc, day post coitum; dpn, day postnatal; wpn, week postnatal.

a degree that differs as a function of species (Tang et al., 2016). In porcine male GCs, persistent DNA methylation is observed in some TEs and in overlapping genes (Gómez-Redondo et al., 2021), suggesting that the lack of DNA methylation erasure in these genes is a safeguard against the mobilization of TEs. In parallel with DNA demethylation, levels of repressive histone marks rise, which prevents the initiation of massive transcriptional activity following genome hypomethylation (Tang et al., 2016). The loss of 5meC and then of repressive histone marks at specific loci, as well as the gain in 5hmC, are all essential for the expression of germline differentiation genes and thus establishment of the germline (Hill et al., 2018).

The DNA re-methylation of GCs uses the de novo methylation enzymes DNMT3A/3B and DNMT3L; the latter is a germline-specific cofactor that is devoid of methyltransferase activity and guides DNMT3A/3B to the sequences to be methylated. In mice, the bulk of de novo DNA methylation occurs during the period of male GC quiescence. In regions associated with euchromatin, the broad deposit of H3K36me2 histone mark, which is recognized and bound by DNMT3A, is necessary for the first wave of de novo DNA methylation (Shirane et al., 2020). In heterochromatin, de novo methylation is delayed and appears to rely on a broad reorganization of chromatin occurring later during mouse development (Yamanaka et al., 2019). PIWI-interacting RNAs contribute to de novo DNA methylation through their role in silencing TEs which try to invade the genome of the germline to be propagated at the next generation. To achieve this silencing, piRNAs displaying partial sequence homology with TEs guide the recruitment of de novo DNA methylation and chromatin remodeling machineries toward nascent TE transcripts (Wang and Lin, 2021). Although most 5meC in male GCs is acquired during life in utero, it appears that modifications still occur after birth (Oakes et al., 2007). Data obtained in mice suggest that the rate and distribution of 5meC in male GCs then stabilize before meiosis, because no important changes can be observed between spermatocytes and spermatozoa (Oakes et al., 2007; Hammoud et al., 2014).

As largely supported by data in mice, the genome-wide erasure and re-apposition of DNA methylation are an important window of epigenetic plasticity that can be altered by deleterious environmental conditions. The living conditions of the mother may affect the reprogramming of male GCs and the sperm methylome in adulthood, with possible physiological effects on reproductive outcomes (Lambrot et al., 2013) and on the metabolism of the next generation (Martínez et al., 2014). In sheep, nutritional stress during pregnancy alters the DNA methylation landscape and the functional parameters of spermatozoa (Toschi et al., 2020). The dynamics of de novo DNA methylation are not yet understood in male cattle; however, the

Downloaded from https://academic.oup.com/af/article/11/6/28/6469044 by INRA Avignon user on 21 September 2022

sperm methylome retains a memory of the nutrition offered during the first months of life (Perrier et al., 2020), suggesting that DNA methylation after birth is still sensitive to environmental factors. Likewise, environmental control during gestation (and particularly the diet of highly producing dams) may prove crucial to ensuring the proper differentiation of male GCs, optimal fertility traits, and an adequate sperm methylome throughout adulthood.

In contrast with the overall stability of DNA methylation during adulthood, chromatin and sncRNA contents are dynamically remodeled during spermatogenesis and beyond. Micro RNAs play an important role in spermatogenesis in mice; they are involved in regulating the differentiation vs. proliferation balance in SSCs (Huang et al., 2017) as well as meiosis and the histone–protamine transition (Liu et al., 2013). Functions in regulation of the stability and translation of mRNAs during mouse spermatogenesis have also been reported for piRNAs, as well as a role in chromosome segregation during meiosis through the regulation of RNAs produced from satellite repeats (Wang and Lin, 2021). Since post-spermiogenesis spermatozoa are transcriptionally inactive, their sncRNAs content was long thought to be stable and exclusively inherited from spermatogenesis. However, it has recently been demonstrated in several species (Chu et al., 2019; Nixon et al., 2019; Sellem et al., 2021) that the sncRNA profile of sperm undergoes important modifications in contact with extracellular vesicles trafficked from epithelial cells in the epididymis (epididymosomes). The piRNA content thus falls markedly during epididymal transit and is replaced by other sncRNA families. In bulls, miRNAs account for 1% of the testicular sperm sncRNA content and then rise to reach 30% in epididymis cauda. The proportion of transfer RNAs- (tRFs) or ribosomal RNAs- (rRFs) derived fragments also increases rapidly as the spermatozoa reach the epididymis (Sellem et al., 2021). The transit of spermatozoa through the epididymis, therefore, represents an important window of epigenetic plasticity, which could be mediated by changes to the sncRNA content of epididymosomes depending on environmental or physiological factors. In line with this view, modifications to the sncRNA profile of sperm have been reported in response to diet in rodents (Grandjean et al., 2015; de Castro Barbosa et al., 2016).

The Epigenome of Bull Sperm and Its Relationships with Fertility

Mature bovine spermatozoa have a particularly low global level of 5meC compared with bovine somatic cells and also to spermatozoa from goats, rams, humans, stallions, boars, and mice. This low 5meC level has been observed in all cattle breeds studied to date and does not seem to be affected by the semen freezing process (Perrier et al., 2018). To determine the undermethylated sequences, the sperm methylome was compared with that of bovine somatic cells using pan-genomic approaches (reduced representation bisulfite sequencing and the immunoprecipitation of methylated DNA followed by hybridization on a microarray). Numerous differentially methylated positions were found, 81% of which were specifically undermethylated in spermatozoa. These undermethylated sites are enriched with spermatogenesis genes and satellite repeats. Overrepresentation of these repeats in the bovine genome, as well as their low methylation, may explain why bovine spermatozoa have lower global 5meC levels than spermatozoa from other mammalian species. Using a different whole-genome approach (whole-genome bisulfite sequencing), another team reported that the DNA methylome of bull sperm contains specific undermethylated domains enriched for satellites and evolutionary young TEs that may escape piRNA-mediated silencing (Zhou et al., 2018). The lower methylation of satellites in sperm compared with somatic cells has been reported in other species (Qu et al., 2018); however, the difference seems to be particularly marked in the bovine species.

The silencing of pericentromeric satellites by the formation of constitutive heterochromatin is essential to maintaining genome stability and preventing both recombination and inappropriate chromosome segregation. In somatic cells, pericentromeric heterochromatin formation is primarily achieved through DNA methylation and the recruitment of heterochromatin protein HP1 to the H3K9me3 histone mark. Some components of the somatic constitutive heterochromatin appear to be maintained in mouse sperm, since satellites escape genome-wide histone–protamine exchange and remain associated with nucleosomes bearing H3K9me3 (Yamaguchi et al., 2018). In addition, histones are detected in distal intergenic regions and CpG-rich promoters and those of developmentally important genes. In bull sperm, two studies have reported partially concordant results regarding the genome-wide location of nucleosomes; interestingly, both agreed to confirm the retention of histones at satellites (Samans et al., 2014; Sillaste et al., 2017).

An exhaustive analysis of the expression profiles of sperm sncRNAs in a cohort of 40 bulls from six breeds was recently carried out (Sellem et al., 2020). Several sncRNA families were detected, including miRNAs (20%), piRNA (26%), rRFs (25%), and tRFs (14%). Interestingly, tRFs associated with glycine or glutamine and derived from the 5′ half of tRNAs were highly represented among all tRNAs. Whatever the sncRNA family, few sequences were predominantly expressed. For instance, the 20 most expressed miRNA sequences accounted for 75% of total miRNA expression, suggesting their functional importance. Numerous isomiRs (sequence variants of canonical miRNAs) were also identified, thus increasing the diversity and complexity of the bull sperm sncRNA repertoire. These variations were not related to the presence of known genetic polymorphisms, suggesting that they could rely on specific RNA edition mechanisms such as the trimming or adding of one or several nucleotides at sequence extremities. Such edition mechanisms have been described in humans, where several nucleotidyl transferases (especially uridyltransferases and adenyltransferases) are involved in the biogenesis of isomiRs (Neilsen et al., 2012). Among all the sequences identified as miRNAs, only 26% have been described and recorded in databases, suggesting that bull sperm contains many novel and

Downloaded from https://academic.oup.com/af/article/11/6/28/6469044 by INRA Avignon user on 21 September 2022

putative miRNAs. Such diversity in the sncRNA content of bull sperm has thus been reported for the first time and was probably determined, thanks to an optimized RNA extraction method and important sequencing depth (Sellem et al., 2020).

Bull semen is a commercial product widely used for AI. Because unsuccessful AI can result in economic losses, extended calving intervals, increased culling rates, and lower rates of genetic gain, several studies (listed below) have investigated the association between interindividual variations of the semen epigenome and fertility traits of bulls. Some studies have highlighted differences in DNA methylation patterns between groups of bulls with different fertility scores (Kropp et al., 2017; Gross et al., 2020; Narud et al., 2021; Takeda et al., 2021). Overall, the results reported were not concordant in terms of the genes or genomic regions targeted by differential methylation and the magnitude of DNA methylation changes. These inconsistencies may be related to both technical issues (e.g., because of the different approaches used to generate DNA methylation data and parameters used to detect differential methylation) and biological issues (different scores used to assess bull fertility, different breeds, high interindividual variability, and small numbers of samples involved in each study). Likewise, several studies have focused on the association between interindividual variations in the sperm sncRNA content and semen quality or bull fertility, highlighting several miRNAs (Capra et al., 2017; Alves et al., 2020; Keles et al., 2021). In addition, other sncRNAs such as tRF-Gly or tRF-Glu may represent another source of fertility biomarkers, as suggested by their differential expression according to in vitro fertilization outcomes in humans (Hua et al., 2019). Due to the high compaction level of sperm chromatin, studies on the genomic location of posttranslational modifications of histones are technically challenging; however, the histone retention degree and associated modifications have been reported to vary as a function of bull fertility using flow cytometry (Ugur et al., 2019).

Because of its multifactorial nature, understanding and predicting fertility are challenging. Furthermore, in the AI industry, routine semen quality control tests are carried out (Figure 1), allowing the identification of most bulls with severe infertility and observable effects on semen functional parameters. Compared with humans, the difference between fertile and subfertile bulls is subtle, hampering the accurate prediction of fertility. For these reasons, most of the studies mentioned above should be regarded as prospective. Larger cohorts of bulls that are well characterized in terms of their genotypes and fertility would be necessary to reduce interindividual variability and to develop models. Integrating various signals (DNA methylation, sncRNAs, and genotypes) rather than considering only one source of information may also provide additional insights into the architecture of fertility. Because spermatozoa are transcriptionally silent, another avenue could arise from functional experiments on the embryo, such as monitoring the effects of the overexpression or suppression of particular miRNAs on the kinetics and quality of embryonic development. This research could have potential applications in human medicine, as AI bulls usually have hundreds of AI records, which considerably alleviates confounding effects and enables the very precise assessment of male fertility.

Contribution of the Paternal Epigenome to Embryonic Development

Evidence demonstrating that the epigenetic information carried by gametes is crucial for development is provided by the poor developmental outcomes of clones (Heyman, 2005). Cloned zygotes are obtained by the transfer of a somatic cell into an enucleated oocyte; they are, therefore, diploid but lack paternal and maternal epigenomes as well as the whole sncRNA content of sperm. The somatic epigenome represents the main barrier to the efficiency of cloning, while the oocyte and sperm epigenomes are extensively reprogrammed after fertilization to allow development of the embryo (Figure 3A).

This reprogramming is characterized by a series of epigenetic modifications that start just after fertilization (Ross and Sampaio, 2018), particularly in the paternal genome. The protamines present on paternal DNA are exchanged with maternal histones, which are rapidly methylated in position H3K4 (activating mark; see Figure 3B, lower panel). By contrast, the maternal chromatin contains numerous repressive histone modifications such as H3K9me3, H3K27me3, and H4K20me3 (Figure 3B, upper panel). This asymmetry between parental genomes eventually fades as embryonic development progresses (Bošković et al., 2012). Overall 5meC levels fall sharply in embryos during preimplantation development (Lepikhov et al., 2008; Dobbs et al., 2013; Bakhtari and Ross, 2014). Genes subject to genomic imprinting are not affected by this wave of DNA demethylation (Jiang et al., 2018; Duan et al., 2019). DNA methylation erasure is more rapid and important in the bovine paternal genome (which is initially more methylated than the maternal genome) and requires the expression of TET enzymes (Bakhtari and Ross, 2014). At the level of the genomic sequence, this overall 5meC decrease is associated with three successive waves of DNA demethylation and de novo methylation. The three steps of DNA methylation erasure coincide with the principal stages of early development and the expression of specific genes, including those encoding de novo DNMTs, which may explain why they are followed by de novo DNA methylation (Jiang et al., 2018).

The epigenetic changes affecting chromatin and DNA methylation participate in triggering embryonic genome activation (**EGA**). Indeed, initially, the genome of the newly fertilized embryo is transcriptionally inactive. Embryo development then depends strictly on the stock of RNA and proteins accumulated in the oocyte (Figure 3A). EGA occurs at the 8-cell stage in cattle and relies on a unique chromatin organization. The overall levels of repressive histone marks reach a minimum level at EGA and recover to the blastocyst stage, as the first cell differentiation occurs. Chromatin accessibility in bovine embryos is also maximal at the time of EGA, and several waves of transcription factor binding sites become accessible from the 2-cell to the morula stages, according to a dynamics that is closer to humans than to mice (Halstead et al., 2020). Likewise,

Downloaded from https://academic.oup.com/af/article/11/6/28/6469044 by INRA Avignon user on 21 September 2022

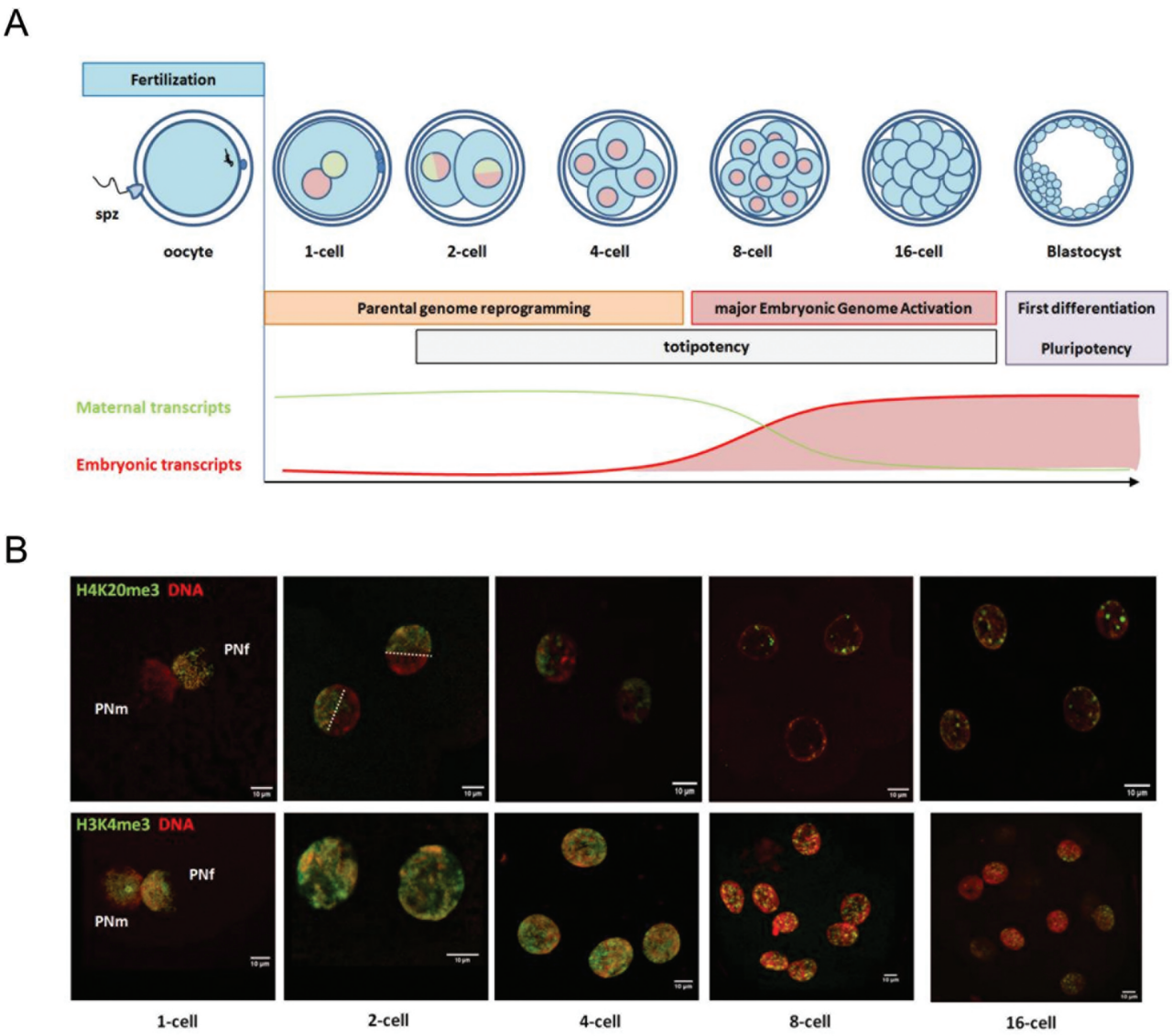


Figure 3. Reprogramming of the parental epigenome during embryonic development in cattle. (A) Upper panel: schematic representation of early embryonic development from fertilization to the first differentiation. Lower panel: dynamics of maternal (green) and embryonic (red) transcripts. (B) Distribution of H4K20me3 (upper panel) and H3K4me3 (lower panel) in the nuclei of bovine embryos from 1-cell to 16-cell stages (unpublished data from our lab). Scale bar: 10 µm. PNf, female pronucleus; PNm, male pronucleus; spz, spermatozoa.

the minimal level of DNA methylation (15%) is coincident with EGA (Duan et al., 2019).

Up to EGA, protein synthesis from the maternal mRNA stock can be regulated by sncRNAs originating from both the oocyte and spermatozoa. These gametic mRNAs and sncRNAs are then progressively diluted as the embryo starts transcribing its own material (Alves et al., 2020). Although the window during which the sperm-borne sncRNAs might exert a regulatory role in the embryo is narrow, they appear to be essential to the normal development of mouse embryos. This was recently illustrated by the developmental arrests and altered transcriptome exhibited by mouse embryos that had been produced using spermatozoa collected in the caput epididymis, which were, therefore, immature regarding their sncRNA content (Conine et al., 2018). Interestingly, the incubation of these immature spermatozoa with cauda epididymosomes restored normal embryonic development, thus demonstrating that essential factors, which may include sncRNAs, are embedded in these epididymosomes.

Potential Mechanisms for the Programming of Phenotypes via the Paternal Route

The sperm epigenetic features transmitted to the embryo are postulated to mediate the intergenerational transmission of nongenetic information that may impact the long-term phenotype of offspring in response to environmental changes affecting the father (Champroux et al., 2018; Donkin and Barrès, 2018). Overall, studies on postfertilization reprogramming in cattle (Duan et al., 2019) and other species have suggested a

Downloaded from https://academic.oup.com/af/article/11/6/28/6469044 by INRA Avignon user on 21 September 2022

limited inheritance of methylated features in the paternal genome. However, individual loci, such as imprinted loci, specific subfamilies of TEs, as well as a few genes, are specifically targeted by the DNA methylation maintenance machinery and are faithfully maintained throughout postfertilization reprogramming in mice (Seah and Messerschmidt, 2018). It has recently been demonstrated in mice that unmethylated CpGs bound by transcription factors expressed in male GCs or in the embryo are protected from de novo DNA methylation, while the absence of transcription factor binding at CpGs already methylated before reprogramming would allow the faithful re-apposition of DNA methylation (Kremsky and Corces, 2020). The authors of this study proposed that any change affecting the expression or binding of these transcription factors during de novo DNA methylation phases may, therefore, stably switch the methylation status of neighboring CpGs, thus offering a novel hypothesis for the mechanistic basis of epigenetic inheritance.

Small noncoding RNAs may also play a role in intergenerational inheritance, as exemplified by studies on the impact of diet on the F0 sncRNA content in rodent sperm. The F1 generation produced with this epigenetically altered sperm is affected by metabolic disorders and displays a modified sperm sncRNA content. Few sncRNAs, such as let-7c, are dysregulated in both F0 and F1 generations. Interestingly, among the let-7c targets, several genes are involved in glucose metabolism and could contribute to the phenotype observed (de Castro Barbosa et al., 2016). Pups developed from zygotes microinjected with the miRNA miR-19b exhibit metabolic alterations similar to those of pups sired by males fed a high-fat diet, suggesting that this particular miRNA instructs the paternal effect induced by diet (Grandjean et al., 2015). Likewise, protein restriction increases the amounts of tRF-Gly in sperm, which in turn modulates transcription in the embryo (Sharma et al., 2016).

Although most histones are removed from the sperm chromatin, those that are retained bear epigenetic marks that could be transmitted to the embryo and mediate intergenerational epigenetic effects (Champroux et al., 2018). In support of this hypothesis, a recent study demonstrated that the distribution of H3K4me3 was altered in the sperm of mice fed a folate-deficient diet from weaning to adulthood. Some of these alterations were also found in the embryos produced using this sperm, which may underlie changes to the post-EGA transcriptome and ultimately lead to developmental defects (Lismer et al., 2021).

Such epigenetic inheritance phenomena via the paternal route have so far not been reported convincingly in cattle but would be of considerable interest in the context of animal selection. Consistent with the major epigenetic reprogramming steps that occur during GC differentiation and after fertilization, modeling approaches have suggested that the overall magnitude of epigenetic inheritance is weak in cattle (Varona et al., 2015). Beyond epigenetic inheritance (Figure 4A), other molecular mechanisms involving the sperm epigenome may also modulate the offspring phenotype. For instance, aberrant epigenetic patterns or sncRNA contents in sperm may interfere with postfertilization reprogramming and alter the timing of the activation of developmentally regulated genes. The resulting embryo may carry subtle molecular, morphological, or metabolic defects that could drive long-term effects on the phenotype, in line with the theory regarding the developmental origin of health and diseases (**DOHaD**, Figure 4B). For instance, the exposure of male GCs to heat stress during spermatogenesis leads to chromatin condensation defects in bull spermatozoa and interferes with the reprogramming of DNA methylation in the paternal pronucleus after fertilization (Rahman et al., 2014). Another example is provided by the epigenetic alterations that have been observed in semen collected at a peripubertal age. Embryos produced using such peripubertal semen display subtle modifications to the DNA methylome and transcriptome that particularly affect the genes involved in metabolic functions and protein synthesis (Wu and Sirard, 2020). It is noteworthy that only morphologically normal embryos, which would likely have implanted and given birth to progeny, were considered during this study. Although later developmental stages were not investigated in these examples, it is possible that epigenetic changes induced by altered preimplantation development might be detected in the offspring. Lastly, some epigenetic features in sperm may be inherited by offspring because they are controlled by genetic mechanisms. Like other genetically controlled phenotypes, these epigenetic features are expected to be heritable. Because selection in cattle is reliant on the association between sire genotypes and daughter performances, it appears particularly challenging to disentangle epigenetic inheritance from genetically controlled epigenetic effects. A clearer understanding of the proportion of the epigenome under genetic control is, therefore, essential if we are to produce an initial estimate of intergenerational epigenetic inheritance in cattle and its impact on phenotypes.

Conclusions

Numerous studies conducted in humans or model species have established the role of in utero conditions in the long-term programming of phenotypes, a phenomenon known as DOHaD. The concept of Paternal Origins of Health and Diseases has emerged more recently (Soubry, 2018), underscoring the importance of the paternal epigenome to offspring phenotype. In cattle, several selection and breeding practices may interfere with proper establishment of the sperm epigenome of bulls used for AI, leading to epigenetic alterations that can potentially disseminate to many herds and have long-term consequences. Windows of epigenome sensitivity to environmental factors exist, during which particular attention should be paid to the bull and its environment in order to maximize the epigenetic potential of sperm. Optimal in utero conditions should first be put in place by monitoring the nutrition, health, and welfare of the dam in order to ensure the proper epigenetic reprogramming and development of fetal GCs. Optimal conditions during the postnatal period, with particular focus on the transition phases, growth and puberty, as well as

Downloaded from https://academic.oup.com/af/article/11/6/28/6469044 by INRA Avignon user on 21 September 2022

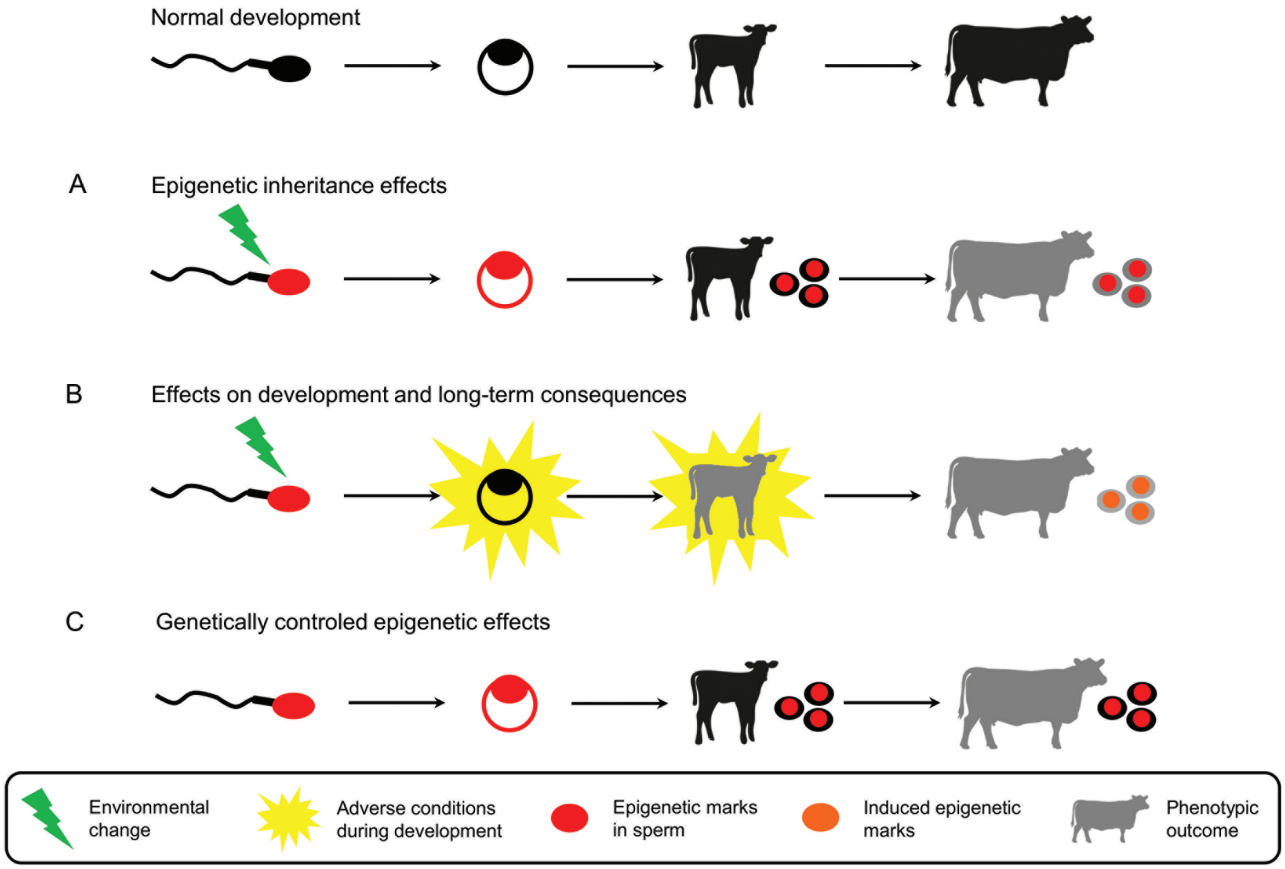


Figure 4. Simplified model for paternal effects on the offspring phenotype in cattle. Environmentally induced epigenetic changes in sperm can either be inherited by the offspring and drive phenotypic changes in the short or longer term (A) or interfere with normal development without transmission (B). The epigenetic marks thus induced and observable in the offspring are then the consequence rather than the cause of adverse conditions during development. (C) The epigenetic marks in sperm can also be transmitted to offspring through genetic mechanisms, leading to phenotypic outcomes independently of environmental changes. These genetically controlled epigenetic effects can easily be confounded with epigenetic inheritance.

during the semen production period, should contribute to early and efficient spermatogenesis, the stable production of high-quality semen, and the overall fertility of the bull. The early culling of AI bulls with poor semen functional parameters is a common practice in the breeding industry, resulting in the elimination of males with severe infertility and thus facilitating the distribution of semen devoid of major epigenetic defects to different herds. On the other hand, the development of embryo biotechnologies such as intra-cytoplasmic sperm injection can to some extent compensate for spermatogenesis defects. As a consequence, bulls with exceptional genetic merit but poor semen quality can now be used in breeding schemes to generate marketed AI bulls whose semen will in turn be distributed extensively to herds. The intergenerational impact of embryo biotechnologies, combined with poor semen quality involving probable epigenetic defects, is still a matter of debate in humans and should also be considered in livestock. The degree to which nongenetic factors carried by sperm can shape the offspring phenotype is an issue that remains largely unsolved in cattle, owing to the delayed collection of performance data relative to gestation, calving, and lactation and to the confounding effect of genetics. The design of affordable epigenotyping tools that could be used on both semen and the blood of daughters, together with the development of integrated approaches that combine both genetic and epigenetic information, may help to address this question.

Acknowledgments

We apologize for all the valuable work that could not be mentioned in this short review due to space constraints. We thank all members of our respective teams, our collaborators, our funders APIS-GENE and agence nationale de la recherche (ANR), as well as the breeding companies such as Évolution, Umotest, Evajura, and AWE. We would also like to thank Victoria Hawken for English language editing.

Literature Cited

Alves, M.B.R., E.C.C. Celeghini, and C. Belleannée. 2020. From sperm motility to sperm-borne microRNA signatures: new approaches to predict male fertility potential. Front. Cell Dev. Biol. 8:791. doi:10.3389/fcell.2020.00791

Aurich, C., B. Schreiner, N. Ille, M. Alvarenga, and D. Scarlet. 2016. Cytosine methylation of sperm DNA in horse semen after cryopreservation. Theriogenology 86:1347–1352. doi:10.1016/j.theriogenology.2016.04.077

Downloaded from https://academic.oup.com/af/article/11/6/28/6469044 by INRA Avignon user on 21 September 2022

About the Authors

Dr. **Hélène Kiefer** obtained a permanent position at INRA (now INRAE) as a research scientist in 2010. She joined a team working on livestock epigenetics and has since developed several tools to study DNA methylation at a genome-wide scale in cattle: a Roche-NimbleGen microarray and more recently RRBS. She collaborates with bioinformatics specialists and statisticians to develop pipelines for DNA methylation analyses. Her research aims to understand how epigenetic modifications vary under the influence of both environmental and genetic factors and contribute to phenotype variability in cattle. She is particularly interested in the bull sperm epigenome and its potential role as a mediator of intergenerational inheritance effects. **Corresponding author:** helene.kiefer@inrae.fr

Dr. Eli Sellem has been working at Allice for 20 yr. He is particularly interested in the enhancement of semen production and semen quality, cryopreservation, and preservation and in predicting the onset of puberty. His key focus is on predicting bull fertility by developing sperm functionality assessments and in the past 7 yr through analysis of the semen epigenome.

Dr. Amélie Bonnet-Garnier is a senior scientist working at the French National Research Institute for Agriculture, Food and Environment (INRAE) in the Biology of Reproduction, Epigenetic, Environment and Development Unit (BREED, UMR 1198, INRAE UVSQ ENVA, UP Saclay). Her work focuses on deciphering how genome organization and epigenetic mechanisms regulate gene expression at the time of embryonic genome activation. She has developed technologies using FISH on 3D-preserved nuclei and image analysis approaches to study the links between large-scale chromatin reorganization and developmental success in several mammalian embryos.

Dr. Maëlle Pannetier obtained her PhD in molecular genetics at the University of Versailles Saint-Quentin, working on sex determination in the goat species at INRAE in Jouy-en-Josas. During her postdoctoral attachment at the Institute of Molecular Genetics in Montpellier (CNRS), she developed her skills and interest in the epigenome, and particularly the histone posttranslational modifications involved in parental imprinting control. She joined INRAE as a research scientist in 2008. She currently manages a small group in the BREED laboratory that works on sex determination and gonad differentiation in ruminant and rabbit species. They are focusing on the differentiation of germ cells and somatic cells from the physiological, genomic, epigenomic, and transcriptomic points of view.

Valentin COSTES is a PhD student from the cooperative union Allice and is working in the host laboratory BREED (Biology of Reproduction, Environment, Epigenetics and Development) at INRAE and Paris-Saclay University. His PhD project consists of the identification of sperm epigenetic biomarkers for different phenotypes in the bovine species, using integrative and modeling approaches.

Dr. Laurent Schibler is currently the head of Allice's Development & Innovation team. After an agricultural engineering degree and a PhD in molecular genetics and structural genomics, he worked for 15 yr at INRA, at the crossroad of several disciplines but mainly on animal genetics and genomics; he thus contributed to the identification of several genes and QTLs of agronomic interest. As a team leader, he has used state-of-the-art genomics and proteomics technologies to study cartilage maturation in several transgenic mouse models and to offer new insights into the molecular mechanisms involved in equine osteochondrosis. His current aim is to develop integrated projects in cattle reproduction, focused in particular on semen fertility and combining genetics, (epi)genomics, and physiology.

Downloaded from https://academic.oup.com/af/article/11/6/28/6469044 by INRA Avignon user on 21 September 2022

Dr. Hélène Jammes is an expert in reproductive physiology. In 1984, she started working at INRAE on the role of gonadotropin in luteal cell activity and growth hormone (GH) transduction signaling in the mammary gland. In 2002, she developed epigenetic analyses of imprinted genes related to human pregnancy pathologies. She has worked on the epigenetic effects induced by Assisted Medical Procreation on embryo and placenta development using mouse models. She has investigated DNA methylation alterations in sperm from infertile men. A little over 10 yr ago, she returned to INRAE and joined UMR 1198 (Biology of Reproduction, Environment, Epigenetics and Development) and currently manages a team of eight permanent staff and students. This team focuses on epigenetic biomarkers for the programming of phenotypes. She also leads studies on DNA methylation as a biomarker for dairy cow health.

Bakhtari, A., and P.J. Ross. 2014. DPPA3 prevents cytosine hydroxymethylation of the maternal pronucleus and is required for normal development in bovine embryos. Epigenetics 9:1271–1279. doi:10.4161/epi.32087

Bošković, A., A. Bender, L. Gall, C. Ziegler-Birling, N. Beaujean, and M.E. Torres-Padilla. 2012. Analysis of active chromatin modifications in early mammalian embryos reveals uncoupling of H2A.Z acetylation and H3K36 trimethylation from embryonic genome activation. Epigenetics 7:747–757. doi:10.4161/epi.20584

Capra, E., F. Turri, B. Lazzari, P. Cremonesi, T.M. Gliozzi, I. Fojadelli, A. Stella, and F. Pizzi. 2017. Small RNA sequencing of cryopreserved semen from single bull revealed altered miRNAs and piRNAs expression between high- and low-motile sperm populations. BMC Genomics 18:14. doi:10.1186/s12864-016-3394-7

Carrell, D.T. 2012. Epigenetics of the male gamete. Fertil. Steril. 97:267–274. doi:10.1016/j.fertnstert.2011.12.036

de Castro Barbosa, T., L.R. Ingerslev, P.S. Alm, S. Versteyhe, J. Massart, M. Rasmussen, I. Donkin, R. Sjögren, J.M. Mudry, L. Vetterli, et al. 2016. High-fat diet reprograms the epigenome of rat spermatozoa and transgenerationally affects metabolism of the offspring. Mol. Metab. 5:184–197. doi:10.1016/j.molmet.2015.12.002

Champroux, A., J. Cocquet, J. Henry-Berger, J.R. Drevet, and A. Kocer. 2018. A decade of exploring the mammalian sperm epigenome: paternal epigenetic and transgenerational inheritance. Front. Cell Dev. Biol. 6:50. doi:10.3389/fcell.2018.00050

Champroux, A., J. Torres-Carreira, P. Gharagozloo, J.R. Drevet, and A. Kocer. 2016. Mammalian sperm nuclear organization: resiliencies and vulnerabilities. Basic Clin. Androl. 26:17. doi:10.1186/s12610-016-0044-5

Chu, C., Y.L. Zhang, L. Yu, S. Sharma, Z.L. Fei, and J.R. Drevet. 2019. Epididymal small non-coding RNA studies: progress over the past decade. Andrology 7:681–689. doi:10.1111/andr.12639

Conine, C.C., F. Sun, L. Song, J.A. Rivera-Pérez, and O.J. Rando. 2018. Small RNAs gained during epididymal transit of sperm are essential for embryonic development in mice. Dev. Cell 46:470–480.e3. doi:10.1016/j.devcel.2018.06.024

Dobbs, K.B., M. Rodriguez, M.J. Sudano, M.S. Ortega, and P.J. Hansen. 2013. Dynamics of DNA methylation during early development of the preimplantation bovine embryo. PLoS One. 8:e66230. doi:10.1371/journal.pone.0066230

Donkin, I., and R. Barrès. 2018. Sperm epigenetics and influence of environmental factors. Mol. Metab. 14:1–11. doi:10.1016/j.molmet.2018.02.006

Duan, J.E., Z.C. Jiang, F. Alqahtani, I. Mandoiu, H. Dong, X. Zheng, S.L. Marjani, J. Chen, and X.C. Tian. 2019. Methylome dynamics of bovine gametes and in vivo early embryos. Front. Genet. 10:512. doi:10.3389/fgene.2019.00512

Gómez-Redondo, I., B. Planells, S. Cánovas, E. Ivanova, G. Kelsey, and A. Gutiérrez-Adán. 2021. Genome-wide DNA methylation dynamics during epigenetic reprogramming in the porcine germline. Clin. Epigenetics 13:27. doi:10.1186/s13148-021-01003-x

Grandjean, V., S. Fourré, D.A. De Abreu, M.A. Derieppe, J.J. Remy, and M. Rassoulzadegan. 2015. RNA-mediated paternal heredity of diet-induced obesity and metabolic disorders. Sci. Rep. 5:18193. doi:10.1038/srep18193

Gross, N., F. Peñagaricano, and H. Khatib. 2020. Integration of whole-genome DNA methylation data with RNA sequencing data to identify markers for bull fertility. Anim. Genet. 51:502–510. doi:10.1111/age.12941

Halstead, M.M., X. Ma, C. Zhou, R.M. Schultz, and P.J. Ross. 2020. Chromatin remodeling in bovine embryos indicates species-specific regulation of genome activation. Nat. Commun. 11:4654. doi:10.1038/s41467-020-18508-3

Hammoud, S.S., D.H. Low, C. Yi, D.T. Carrell, E. Guccione, and B.R. Cairns. 2014. Chromatin and transcription transitions of mammalian adult germline stem cells and spermatogenesis. Cell Stem Cell 15:239–253. doi:10.1016/j.stem.2014.04.006

Heyman, Y. 2005. Nuclear transfer: a new tool for reproductive biotechnology in cattle. Reprod. Nutr. Dev. 45:353–361. doi:10.1051/rnd:2005026

Hill, P.W.S., H.G. Leitch, C.E. Requena, Z. Sun, R. Amouroux, M. Roman-Trufero, M. Borkowska, J. Terragni, R. Vaisvila, S. Linnett, et al. 2018. Epigenetic reprogramming enables the transition from primordial germ cell to gonocyte. Nature 555:392–396. doi:10.1038/nature25964

Hua, M., W. Liu, Y. Chen, F. Zhang, B. Xu, S. Liu, G. Chen, H. Shi, and L. Wu. 2019. Identification of small non-coding RNAs as sperm quality biomarkers for in vitro fertilization. Cell Discov. 5:20. doi:10.1038/s41421-019-0087-9

Huang, Y.L., G.Y. Huang, J. Lv, L.N. Pan, X. Luo, and J. Shen. 2017. miR-100 promotes the proliferation of spermatogonial stem cells via regulating Stat3. Mol. Reprod. Dev. 84:693–701. doi:10.1002/mrd.22843

Jiang, Z., J. Lin, H. Dong, X. Zheng, S.L. Marjani, J. Duan, Z. Ouyang, J. Chen, and X.C. Tian. 2018. DNA methylomes of bovine gametes and in vivo produced preimplantation embryos. Biol. Reprod. 99:949–959. doi:10.1093/biolre/ioy138

Keles, E., E. Malama, S. Bozukova, M. Siuda, S. Wyck, U. Witschi, S. Bauersachs, and H. Bollwein. 2021. The micro-RNA content of unsorted cryopreserved bovine sperm and its relation to the fertility of sperm after sex-sorting. BMC Genomics 22:30. doi:10.1186/s12864-020-07280-9

Kremsky, I., and V.G. Corces. 2020. Protection from DNA re-methylation by transcription factors in primordial germ cells and pre-implantation embryos can explain trans-generational epigenetic inheritance. Genome Biol. 21:118. doi:10.1186/s13059-020-02036-w

Kropp, J., J.A. Carrillo, H. Namous, A. Daniels, S.M. Salih, J. Song, and H. Khatib. 2017. Male fertility status is associated with DNA methylation signatures in sperm and transcriptomic profiles of bovine preimplantation embryos. BMC Genomics 18:280. doi:10.1186/s12864-017-3673-y

Lambrot, R., C. Xu, S. Saint-Phar, G. Chountalos, T. Cohen, M. Paquet, M. Suderman, M. Hallett, and S. Kimmins. 2013. Low paternal dietary folate alters the mouse sperm epigenome and is associated with negative pregnancy outcomes. Nat. Commun. 4:2889. doi:10.1038/ncomms3889

Lepikhov, K., V. Zakhartchenko, R. Hao, F. Yang, C. Wrenzycki, H. Niemann, E. Wolf, and J. Walter. 2008. Evidence for conserved DNA and histone H3 methylation reprogramming in mouse, bovine and rabbit zygotes. Epigenetics Chromatin 1:8. doi:10.1186/1756-8935-1-8

Lismer, A., V. Dumeaux, C. Lafleur, R. Lambrot, J. Brind´Amour, M.C. Lorincz, and S. Kimmins. 2021. Histone H3 lysine 4 trimethylation in sperm is transmitted to the embryo and associated with diet-induced phenotypes in the offspring. Dev. Cell 56:671–686.e6. doi:10.1016/j.devcel.2021.01.014

Liu, T., Y. Huang, J. Liu, Y. Zhao, L. Jiang, Q. Huang, W. Cheng, and L. Guo. 2013. MicroRNA-122 influences the development of sperm abnormalities

Downloaded from https://academic.oup.com/af/article/11/6/28/6469044 by INRA Avignon user on 21 September 2022

from human induced pluripotent stem cells by regulating TNP2 expression. Stem Cells Dev. 22:1839–1850. doi:10.1089/scd.2012.0653

Martínez, D., T. Pentinat, S. Ribó, C. Daviaud, V.W. Bloks, J. Cebrià, N. Villalmanzo, S.G. Kalko, M. Ramón-Krauel, R. Díaz, et al. 2014. In utero undernutrition in male mice programs liver lipid metabolism in the second-generation offspring involving altered Lxra DNA methylation. Cell Metab. 19:941–951. doi:10.1016/j.cmet.2014.03.026

Narud, B., A. Khezri, T.T. Zeremichael, E.B. Stenseth, B. Heringstad, A. Johannisson, J.M. Morrell, P. Collas, F.D. Myromslien, and E. Kommisrud. 2021. Sperm chromatin integrity and DNA methylation in Norwegian Red bulls of contrasting fertility. Mol. Reprod. Dev. 88:187–200. doi:10.1002/mrd.23461

Neilsen, C.T., G.J. Goodall, and C.P. Bracken. 2012. IsomiRs—the overlooked repertoire in the dynamic microRNAome. Trends Genet. 28:544–549. doi:10.1016/j.tig.2012.07.005

Nixon, B., G.N. De Iuliis, M.D. Dun, W. Zhou, N.A. Trigg, and A.L. Eamens. 2019. Profiling of epididymal small non-protein-coding RNAs. Andrology 7:669–680. doi:10.1111/andr.12640

Oakes, C.C., S. La Salle, D.J. Smiraglia, B. Robaire, and J.M. Trasler. 2007. Developmental acquisition of genome-wide DNA methylation occurs prior to meiosis in male germ cells. Dev. Biol. 307:368–379. doi:10.1016/j.ydbio.2007.05.002

Perrier, J.P., D.A. Kenny, A. Chaulot-Talmon, C.J. Byrne, E. Sellem, L. Jouneau, A. Aubert-Frambourg, L. Schibler, H. Jammes, P. Lonergan, et al. 2020. Accelerating onset of puberty through modification of early life nutrition induces modest but persistent changes in bull sperm DNA methylation profiles post-puberty. Front. Genet. 11:945. doi:10.3389/fgene.2020.00945

Perrier, J.P., E. Sellem, A. Prézelin, M. Gasselin, L. Jouneau, F. Piumi, H. Al Adhami, M. Weber, S. Fritz, D. Boichard, et al. 2018. A multi-scale analysis of bull sperm methylome revealed both species peculiarities and conserved tissue-specific features. BMC Genomics 19:404. doi:10.1186/s12864-018-4764-0

Qu, J., E. Hodges, A. Molaro, P. Gagneux, M.D. Dean, G.J. Hannon, and A.D. Smith. 2018. Evolutionary expansion of DNA hypomethylation in the mammalian germline genome. Genome Res. 28:145–158. doi:10.1101/gr.225896.117

Rahman, M.B., M.M. Kamal, T. Rijsselaere, L. Vandaele, M. Shamsuddin, and A. Van Soom. 2014. Altered chromatin condensation of heat-stressed spermatozoa perturbs the dynamics of DNA methylation reprogramming in the paternal genome after in vitro fertilisation in cattle. Reprod. Fertil. Dev. 26:1107–1116. doi:10.1071/RD13218

Rivera, R.M. 2019. Consequences of assisted reproductive techniques on the embryonic epigenome in cattle. Reprod. Fertil. Dev. 32:65–81. doi:10.1071/RD19276

Ross, P.J., and Sampaio, R. V. 2018. Epigenetic remodeling in preimplantation embryos: cows are not big mice in 32nd Annual Meeting of the Brazilian Embryo Technology Society (SBTE) Vol. 15 (ed Animal Reproduction) 204-2014 (Florianopólis, Brazil, 2018). doi:10.21451/1984-3143-AR2018-0068

Samans, B., Y. Yang, S. Krebs, G.V. Sarode, H. Blum, M. Reichenbach, E. Wolf, K. Steger, T. Dansranjavin, and U. Schagdarsurengin. 2014. Uniformity of nucleosome preservation pattern in mammalian sperm and its connection to repetitive DNA elements. Dev. Cell 30:23–35. doi:10.1016/j.devcel.2014.05.023

Seah, M.K.Y., and D.M. Messerschmidt. 2018. From germline to soma: epigenetic dynamics in the mouse preimplantation embryo. Curr. Top. Dev. Biol. 128:203–235. doi:10.1016/bs.ctdb.2017.10.011

Sellem, E., S. Marthey, A. Rau, L. Jouneau, A. Bonnet, C. Le Danvic, B. Guyonnet, H. Kiefer, H. Jammes, and L. Schibler. 2021. Dynamics of cattle sperm sncRNAs during maturation, from testis to ejaculated sperm. Epigenetics Chromatin 14:24. doi:10.1186/s13072-021-00397-5

Sellem, E., S. Marthey, A. Rau, L. Jouneau, A. Bonnet, J.P. Perrier, S. Fritz, C. Le Danvic, M. Boussaha, H. Kiefer, et al. 2020. A comprehensive overview of bull sperm-borne small non-coding RNAs and their diversity across breeds. Epigenetics Chromatin 13:19. doi:10.1186/s13072-020-00340-0

Sharma, U., C.C. Conine, J.M. Shea, A. Boskovic, A.G. Derr, X.Y. Bing, C. Belleannee, A. Kucukural, R.W. Serra, F. Sun, et al. 2016. Biogenesis and function of tRNA fragments during sperm maturation and fertilization in mammals. Science 351:391–396. doi:10.1126/science.aad6780

Shirane, K., F. Miura, T. Ito, and M.C. Lorincz. 2020. NSD1-deposited H3K36me2 directs de novo methylation in the mouse male germline and counteracts polycomb-associated silencing. Nat. Genet. 52:1088–1098. doi:10.1038/s41588-020-0689-z

Sillaste, G., L. Kaplinski, R. Meier, Ü. Jaakma, E. Eriste, and A. Salumets. 2017. A novel hypothesis for histone-to-protamine transition in *Bos taurus* spermatozoa. Reproduction 153:241–251. doi:10.1530/REP-16-0441

Soubry, A. 2018. POHaD: why we should study future fathers. Environ. Epigenet. 4:1–7. doi:10.1093/eep/dvy007

Staub, C., and L. Johnson. 2018. Review: Spermatogenesis in the bull. Animal 12(s1):s27–s35. doi:10.1017/S1751731118000435

Takeda, K., E. Kobayashi, K. Ogata, A. Imai, S. Sato, H. Adachi, Y. Hoshino, K. Nishino, M. Inoue, M. Kaneda, et al. 2021. Differentially methylated CpG sites related to fertility in Japanese Black bull spermatozoa: epigenetic biomarker candidates to predict sire conception rate. J. Reprod. Dev. 67:99–107. doi:10.1262/jrd.2020-137

Tang, W.W., T. Kobayashi, N. Irie, S. Dietmann, and M.A. Surani. 2016. Specification and epigenetic programming of the human germ line. Nat. Rev. Genet. 17:585–600. doi:10.1038/nrg.2016.88

Toschi, P., E. Capra, D.A. Anzalone, B. Lazzari, F. Turri, F. Pizzi, P.A. Scapolo, A. Stella, J.L. Williams, P. Ajmone Marsan, et al. 2020. Maternal periconceptional undernourishment perturbs offspring sperm methylome. Reproduction 159:513–523. doi:10.1530/REP-19-0549

Ugur, M.R., N.A. Kutchy, E.B. de Menezes, A. Ul-Husna, B.P. Haynes, A. Uzun, A. Kaya, E. Topper, A. Moura, and E. Memili. 2019. Retained acetylated histone four in bull sperm associated with fertility. Front. Vet. Sci. 6:223. doi:10.3389/fvets.2019.00223

Varona, L., S. Munilla, E.F. Mouresan, A. González-Rodríguez, C. Moreno, and J. Altarriba. 2015. A Bayesian model for the analysis of transgenerational epigenetic variation. G3 (Bethesda). 5:477–485. doi:10.1534/g3.115.016725

Wang, C., and Lin, H. 2021. Roles of piRNAs in transposon and pseudogene regulation of germline mRNAs and lncRNAs. Genome Biol. 22:27. doi:10.1186/s13059-020-02221-x

Wrobel, K.H. 2000. Prespermatogenesis and spermatogoniogenesis in the bovine testis. Anat. Embryol. (Berl). 202:209–222. doi:10.1007/s004290000111

Wu, C., and M.A. Sirard. 2020. Parental effects on epigenetic programming in gametes and embryos of dairy cows. Front. Genet. 11:557846. doi:10.3389/fgene.2020.557846

Yamaguchi, K., M. Hada, Y. Fukuda, E. Inoue, Y. Makino, Y. Katou, K. Shirahige, and Y. Okada. 2018. Re-evaluating the localization of sperm-retained histones revealed the modification-dependent accumulation in specific genome regions. Cell Rep. 23:3920–3932. doi:10.1016/j.celrep.2018.05.094

Yamanaka, S., H. Nishihara, H. Toh, L.A. Eijy Nagai, K. Hashimoto, S.J. Park, A. Shibuya, A.M. Suzuki, Y. Tanaka, K. Nakai, et al. 2019. Broad heterochromatic domains open in gonocyte development prior to de novo DNA methylation. Dev. Cell 51:21–34.e5. doi:10.1016/j.devcel.2019.07.023

Zhou, Y., Connor, E.E., D.M. Bickhart, C. Li, R.L. Baldwin, S.G. Schroeder, B.D. Rosen, L. Yang, C.P. Van Tassell, and G.E. Liu. 2018. Comparative whole genome DNA methylation profiling of cattle sperm and somatic tissues reveals striking hypomethylated patterns in sperm. GigaScience 7:1–13. doi:10.1093/gigascience/giy039

Downloaded from https://academic.oup.com/af/article/11/6/28/6469044 by INRA Avignon user on 21 September 2022

Communications

Articles scientifiques

Predicting male fertility from the sperm methylome : application to 120 bulls with hundreds of artificial insemination records. Valentin Costes, Aurélie Chaulot-Talmon, Eli Sellem, Jean-Philippe Perrie, Anne Aubert-Frambourg, Luc Jouneau, Charline Pontlevoy, Chris Hozé, Sébastien Fritz, Mekki Boussaha, Chrystelle Le Danvic, Marie-Pierre Sanchez, Didier Boichard, Laurent Schibler, Hélène Jammes, Florence Jaffrézic et Hélène Kiefer. Clinical Epigenetics. Sous presse.

Integrative analysis of semen parameters, genotypes, DNA methylation and sncRNAs from spermatozoa to predict bull fertility. Valentin Costes, Eli Sellem, Sylvain Marthey, Chris Hoze, Aurélie Bonnet, Laurent Schibler, Hélène Kiefer and Florence Jaffrezic. En relecture pas les co-auteurs.

The epigenome of male germ cells ans the programming of phenotypes in cattle. Hélène Kiefer, Eli Sellem, Aurélie Bonnet-Garnier, Maëlle Pannetier, Valentin Costes, Laurent Schibler et Hélène Jammes. Animal Frontiers. 2021.

Communication affichée

Communication internationale :
Prediction of bull fertility based on the sperm methylome. DADE (Domestic Animals DOHaD and Epigenetics), 2021.

Communications orales

Communication internationale :
Machine learning and epigenetics : prediction of bull fertility based on sperm methylome. 72ème meeting annule de l'EAAP, Davos, 2021.

Communications réseau national
Apprentissage automatique et épigénétique : prédiction de la fertilité des taureaux à partir du méthylome spermatique. 6ème journée d'animation EpiPHASE, 2021.

Analyse du méthylome spermatique dans le but de prédire la fertilité de taureaux. CATI INRAE, 2021.

Communication invitée
Intégration de données –omiques pour prédire la fertilité des taureaux. Séminaire interne de l'équipe du Dr Vaiman, 2022.

Autres communications
Séminaire unité GABI

Journée des doctorants de l'unité GABI (2ème année)

Journée des doctorants de l'unité BREED (1ère, 2ème et 3ème année)