TOWARDS IMPROVEMENT OF **RUMI**NANT BREEDING
THROUGH **GEN**OMIC AND EPIGENOMIC APPROACHES

Grant agreement number: 101000226

H2020 – Research and Innovation Action

Deliverable 6.3

Report on the genetic control of DNA methylation

Due date of deliverable: M43 – December 2024

Actual submission date: M44 – January 2024

WP concerned	WP6	
Lead partner	Marie-Pierre SANCHEZ	INRAE
Author(s)	Marie-Pierre SANCHEZ Corentin FOUERE Chris HOZE Mekki BOUSSAHA Didier BOICHARD Hélène KIEFER Gabriel COSTA MONTEIRO MOREIRA Amrita RAJA-RAVI-SHANKAR Aurélie CHAULOT-TALMON Francesca ALI Valentin COSTES Chrystelle LE DANVIC	INRAE Eliance Eliance INRAE INRAE INRAE INRAE INRAE INRAE INRAE Eliance Eliance
Version	V1	
Dissemination level	CO: Confidential, only for members of the consortium (including the Commission Services)	

Call: H2020-SFS-2020-2

Topic: SFS-13-2020 ● Genome and epigenome enabled breeding in terrestrial livestock

Start date of the project: June 1st, 2021 **Duration:** 60 months

End date of the project: May 31st, 2026 **Project ID:** 101000226

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°101000226.

Table of content

1 Summary

The D6.3 deliverable aims to elucidate the genetic control of DNA methylation through a comprehensive GWAS study to identify loci controlling methylation levels (meQTL) in bull sperm. This work is a collaborative effort between the BREED and GABI units from INRAE (France) and ELIANCE (France), as part of task 6.3, titled “Deciphering the genetic control of DNA methylation”. Analyses have been conducted using data from sperm collected from 405 Holstein bulls.

From about 1.2 million methylation marks characterized, 166,985 base-level DNA methylation measures of CpGs were variable across bulls and have been selected for genetic analyses. CpG colocalizing with genetic polymorphisms were excluded. The heritability of these phenotypes was estimated with a genomic relationship matrix built from SNP genotypes. Methylation QTL were identified by GWAS at the whole sequence level for approximately 13 million SNPs obtained through imputation.

Out of the 166,985 methylation sites analyzed, approximately 127,000 exhibited heritability (h^2) estimates higher than 0.1. Across all CpGs sites, the mean h^2 estimate was 0.26, indicating that sperm methylation rates are under genetic control. A total of 36.5% of these CpGs were significantly associated with at least one SNP. The majority of these SNPs (90.2%) were cis-meQTLs (defined as SNPs located within 1 Mb upstream or downstream of the associated CpG they regulate) while only 9.8% were trans-acting meQTLs (defined as SNPs located greater than 2 Mb of the associated CpG on the same chromosome or if the variant and the CpG were located on different chromosomes). By combining the results of distant associations, it appears that several SNPs affected the methylation of multiple CpGs in trans. These SNPs co-localizing with the *HDAC1*, *KAT2B*, *PRDM9*, *RBBP4*, *SPOCD1*, *BANP*, *TCL1B*, *TCL1A*, *PRKCB*, and *NAP1L4* genes, which are promising functional candidates for regulating methylation.

These results pave the way for deciphering the genetic mechanisms controlling DNA methylation in cattle.

2 Introduction

DNA methylation, an essential epigenetic modification, involves the addition of a methyl group to the cytosine residues of DNA. This process plays a key role in regulating gene expression, maintaining genome stability, and is essential for normal development. Although DNA methylation patterns can be influenced by environmental factors, recent studies have shown that a significant proportion of these patterns is under genetic control and is heritable. This genetic control of DNA methylation must be distinguished from intergenerational epigenetic inheritance, which involves methylation marks that escape the post-fertilization erasure and are thus transmitted to next generation without changes to the underlying DNA sequence.

Understanding the genetic basis of DNA methylation is essential for elucidating its role in complex traits and diseases. In livestock, such as cattle, DNA methylation patterns in sperm are of particular interest because they can influence offspring development and performance traits. Identifying genetic loci that control methylation levels (meQTLs) can provide insights into the mechanisms regulating these epigenetic modifications and offer potential markers for improving breeding programs.

One of the main objectives of task 6.3 was to perform a comprehensive genetic analysis to estimate the heritability of DNA methylation and to conduct a genome-wide association study (GWAS) to pinpoint loci controlling methylation levels (meQTL) in sperm of Holstein bulls.

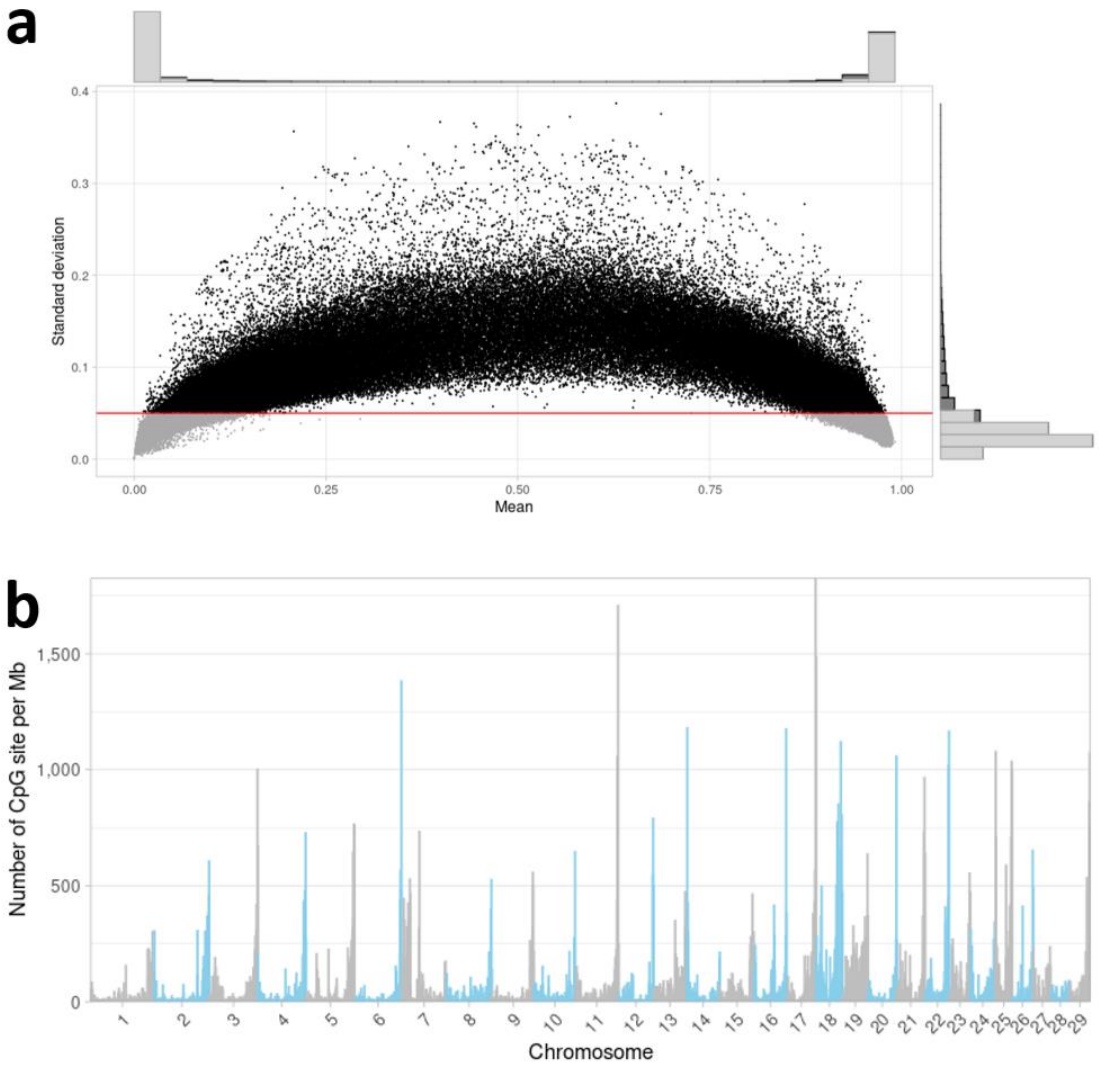
3 Material and Methods

1.1. Data

The sperm genomic DNA from 405 Holstein bulls was extracted in the presence of DTT and proteinase K and analyzed using Reduced Representation Bisulfite Sequencing (RRBS, Integragen SA), using the conditions described in Costes et al., 2022. High quality sequences were then aligned using Bismark (Krueger and Andrews, 2011) on an *in silico* bisulfite converted reference genome (ARS-UCD1.2). DNA methylation rates were obtained from uniquely mapped sequences covering 2 million CpG sites with at least 10X depth. As shown in Figure 1, the majority of CpG sites were either fully unmethylated or fully methylated, making them non-informative for genetic analyses. Therefore, we selected CpG sites with methylation standard deviation above 5%, with less than 20% missing values, and those not co-localized with known SNPs. After filtering, 166,985 individual CpG sites were retained for further analysis.

In addition to methylation data, the bulls were genotyped using the Illumina EuroGMD BeadChip, which contains 53,469 autosomal SNPs. These SNPs passed stringent quality control filters, including an individual call rate greater than 95%, a SNP call rate greater than 90%, a minor allele frequency (MAF) greater than 2%, and genotype frequencies in Hardy-Weinberg equilibrium with $P > 10^{-4}$.

Figure 1: DNA methylation measured by RRBS.



a Scatter plot of CpG methylation levels, showing the mean methylation ratio (x-axis) vs the standard deviation (s.d.) (y-axis) for all CpGs (N = 1,212,361). Black dots, above the red line (s.d. > 0.05), were retained for the study (N = 166,685). The upper and right-sided histograms represent the overall distribution of methylation levels and standard deviations, with retained CpGs highlighted in dark grey. **b** Distribution of the 166,685 selected CpGs per Mb across the genome, with alternating colors distinguishing chromosomes.

1.2. Sequence-based imputation and GWAS

50k genotypes were imputed at the whole genome sequence level using a stepwise approach: 1) high density (777k) genotypes were imputed with FImpute (Sargolzaei et al., 2014) using 804 Holstein major ancestors as a reference, and 2) sequence variants were imputed using 2712 multi-breed animals from the RUN9 reference panel of the 1000 Bull Genomes consortium (Daetwyler et al., 2014), including 1059 Holstein, with Minimac software (Howie et al., 2012).

After filtering, 12,930,389 biallelic variants were tested for association with different DNA methylation traits in each population in separate analyses using GCTA software (Yang et al., 2011), accounting for a polygenic effect with a genomic relationship matrix estimated from 50k SNPs. The following linear mixed model was applied for each of the 166,985 methylation sites:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{x}b + \mathbf{g} + \mathbf{e}, \quad (1)$$

where **y** is the vector of phenotypes (individual methylation rates); μ is the overall mean; b is the additive fixed effect of the variant tested; **x** is the vector of imputed allele dosages for the tested variant; **g ~** N(**0**, **G** σ^2_g) is the vector of random polygenic effects, with **G** the genomic relationship matrix based on 50k autosomal SNPs, and σ^2_g the polygenic variance; and **e ~** N(**0**, **I** σ^2_e) is the vector of random residual effects, with σ^2_e the residual variance and **I** the identity matrix.

Heritability (h^2) estimates were obtained using the following formula:

$$\mathbf{h^2} = \sigma^2_g/(\sigma^2_g+\sigma^2_e) \quad (2)$$

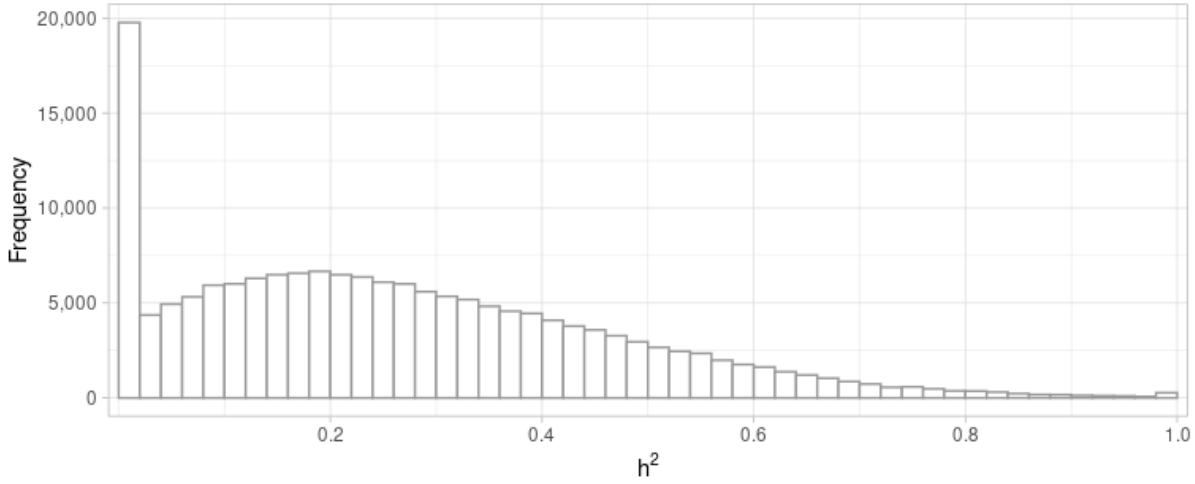
We applied permutation testing to determine significant thresholds for both *cis*- and *trans*-meQTL. For each permutation, genotypes were randomly reassigned to bulls while maintaining the original EuroGMD SNP-based GRM. This approach accounted for population structure while disrupting the link between genotype and phenotype. To reduce the computational burden, GWASs were performed for 10% of CpGs (randomly selected) with permuted genotypes, and the lowest P-value (for *trans*-meQTLs), as well as the lowest P-value within 1 Mb (for *cis*-meQTL), were retained to approximate the distributions of P-values under the null hypothesis. This process was repeated 10 times. The resulted significance thresholds were $P < 1.38 \times 10^{-4}$ for *cis*-meQTLs and $P < 1.25 \times 10^{-8}$ for *trans*-meQTLs, which corresponded to an empirical false discovery rate (FDR) of 5%.

4 Results

4.1. Heritability of methylation rates

Out of the 166,985 methylation sites analyzed, approximately 127,000 exhibited heritability estimates higher than 0.1 (Figure 2). The mean h^2 value was 0.26, indicating that sperm DNA methylation is under genetic control for a large share of the analysed loci.

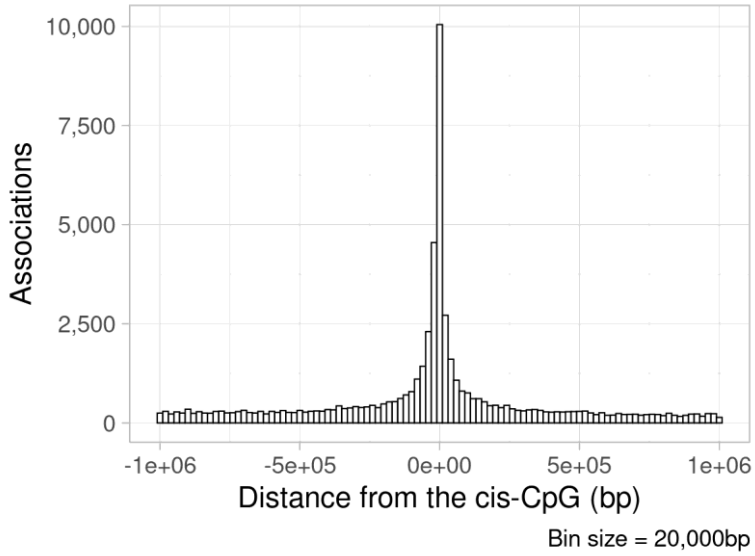
Figure 2: Distribution of heritability estimates for methylation rates



4.2. Identification of meQTL

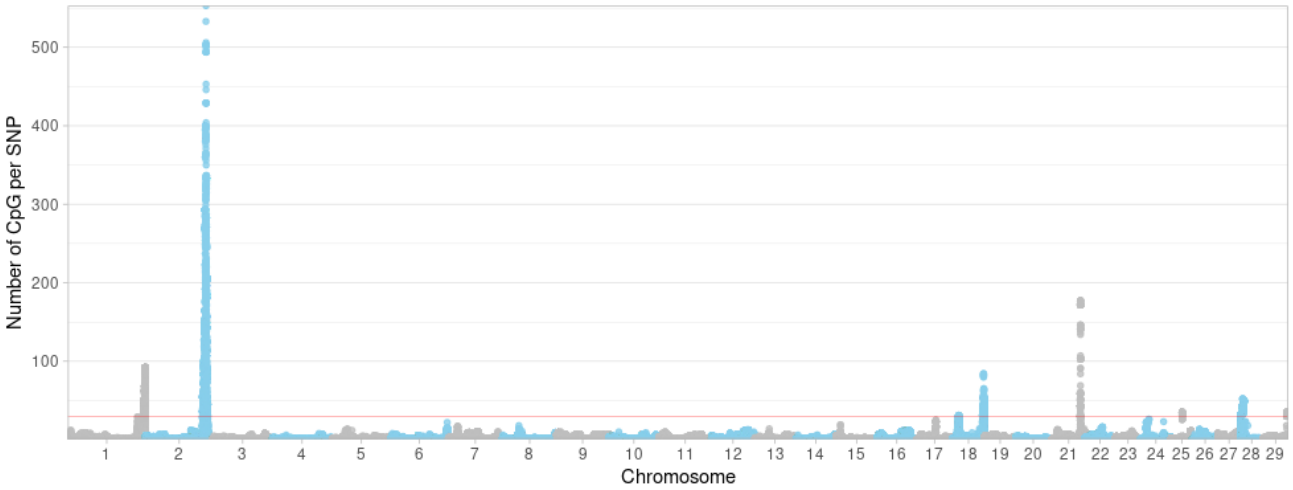
A total of 54,934 CpG sites (32.9%) were associated with at least one SNP. The majority of these SNPs (90.2%) with a significant effect on methylation rates were cis-meQTLs. Most of the cis-meQTLs were found in close proximity to the CpG sites within 100 kb, and the mean absolute distance was 260.8 kb due to the long-range linkage disequilibrium and heavy tails in the distribution (Figure 3).

Figure 3: Distribution of the distance between CpG and SNP for cis-meQTL.



Trans-meQTLs were less numerous, representing 9.8% of all meQTL identified. Some trans-meQTL were associated with the methylation of multiple CpG sites. The SNP that influenced the methylation of the highest number of CpG sites was located on chromosome 2 at 122,297,963 bp. This SNP was associated with the methylation of 553 CpGs (Figure 4) distributed across the entire genome and was located in the vicinity of the *HDAC1* gene which encodes *histone deacetylase 1*. Two other SNPs affecting 177 and 92 CpG sites co-localized with *KAT2B* and *PRDM9* (chromosome 1) and *TLC1B and TCL1A* (chromosome 21), respectively. Given the interplay between DNA methylation and histone post-translational modifications, *HDAC1*, *KAT2B*, *PRDM9*, *RBBP4*, *SPOCD1*, *BANP*, *TCL1B*, *TCL1A*, *PRKCB*, and *NAP1L4* are promising functional candidates for regulating methylation. Interestingly, several of these genes are involved in the deep chromatin changes that occurs during spermatogenesis (Barbero et al., 2023).

Figure 4: Number of CpGs associated with the SNPs of the trans-meQTL



Enrichment was then examined in several genomic annotations, distinguishing cis-CpGs, trans-CpGs, top cis-SNPs (based on the lowest p-value for each cis-CpG) and top trans-SNPs. Overall, the enrichment was moderate for both cis- and trans-CpGs, with odds ratios (OR) ranging from 0.76 to 1.34. Regarding the CpG landscape, both cis- and trans-SNPs were enriched for CpG-dense regions (CpG islands). These two categories were also enriched for several functional genomic annotations including transcription start site (TSS) surrounding regions (TSS200 and TSS150) and in ATAC-seq peaks (retrieved from Yuan et al., 2023), indicators of open chromatin regions.

4.4. Removal of spurious trans-associations

Among the distant associations, a specific pattern was observed for several regions: a group of CpGs was strongly associated with a group of SNPs, and these associations were almost exclusive, meaning that the SNPs were only associated with the CpGs of the distant region. We hypothesised that these associations might be due to a bias in the use of the reference genome assembly (ARS-UCD1.2). To test this hypothesis, we aligned the CpG regions to another publicly available genome assembly (accession number = GCA_021347905.1). For most regions, the alignment differed from the original one observed on the reference genome, and overlapped with the trans-associated region, thus indicating that these observations were likely to be attributed to a small translocation and therefore to cis-meQTL rather than trans-meQTLs.

5 Conclusion

In conclusion, the identification of meQTLs is the subject of intense research to gain a better understanding of the links between genes and phenotypes. To this end, we mapped both *cis*- and *trans*-meQTLs in bull semen. Our results support the hypothesis that the influence of genetic variants on multiple CpGs may be mediated through changes in chromatin structure, possibly involving histone modifiers genes. These results open interesting scientific perspectives, as intense and specific chromatin remodelling takes place during spermatogenesis to allow the compaction of sperm nucleus and the protection of sperm DNA against oxidation. These findings further enhance our understanding of the genetic architecture of bovine semen methylation, a yet unexplored field.

6 References

Barbero, G., M. G. de Sousa Serro, C. Perez Lujan, A. D. Vitullo, C. R. Gonzalez, and B. Gonzalez. 2023. Transcriptome profiling of histone writers/erasers enzymes across spermatogenesis, mature sperm and pre-cleavage embryo: Implications in paternal epigenome transitions and inheritance mechanisms. Frontiers in cell and developmental biology 11:1086573.

Costes V., Chaulot-Talmon A., Sellem E., Perrier J.-P., Aubert-Frambourg A., Jouneau L., Pontlevoy C., Hozé C, Fritz S., Boussaha M., Le Danvic C., Sanchez M-P., Boichard D., Schibler L., Jammes H., Jaffrezic F., Kiefer H. 2022. Predicting male fertility from the sperm methylome: application to 120 bulls with hundreds of artificial insemination records. Clinical Epigenetics, 14:54. https://doi.org/10.1186/s13148-022-01275-x

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. Van Binsbergen, R. F. Brøndum, X. Liao, A. Djari, S. Rodriguez, C. Grohs, D. Esquerré, O. Bouchez, M. N. Rossignol, C. Klopp, D. Rocha, S. Fritz, A. Eggen, P. J. Bowman, D. Coote, A. J. Chamberlain, C. P. VanTassell, I. Hulsegge, M. E. Goddard, B. Guldbrandtsen, M. S. Lund, R. F. Veerkamp, D. Boichard, R. Fries, and B. J. Hayes. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet 46:858-867.

Howie, B., C. Fuchsberger, M. Stephens, J. Marchini, and G. R. Abecasis. 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet 44.

Krueger, F. and S. R. Andrews. 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27(11):1571-1572.

Sargolzaei, M., J. Chesnais, and F. Schenkel. 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genom 15:478.

Yang, J., S. Lee, M. Goddard, and P. Visscher. 2011. GCTA: A Tool for Genome-wide Complex Trait Analysis. Am J Hum Genet 88:76-82.

Yuan C, Tang L, Lopdell T, Petrov VA, Oget-Ebrad C, Moreira GCM, Gualdrón Duarte JL, Sartelet A, Cheng Z, Salavati M, Wathes DC, Crowe MA; GplusE Consortium; Coppieters W, Littlejohn M, Charlier C, Druet T, Georges M, Takeda H. An organism-wide ATAC-seq peak catalog for the bovine and its use to identify regulatory variants. Genome Res. 2023 Oct;33(10):1848-1864. doi: 10.1101/gr.277947.123.